An Examination of the Difference in Reading and Mathematics Achievement between Black and White Students in 1st, 3rd, and 5th Grades and its Relationship to Family, School, and Classroom Variables

By

Ray D. Niboro

Submitted to the graduate degree program in Psychology and Research in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

____

Chairperson Bruce B. Frey, Ph.D.

____

Vicki Peyton, Ph.D.

____

Michael (Mickey) Imber, Ph.D.

____

David M. Hansen, Ph.D.

____

Meagan M. Patterson, Ph.D.

Date Defended: May 6, 2015

The Dissertation Committee for Ray D. Niboro

certifies that this is the approved version of the following dissertation:

An Examination of the Difference in Reading and Mathematics Achievement between Black and White Students in 1st, 3rd, and 5th Grades and its Relationship to Family, School, and Classroom Variables

Chairperson Bruce B. Frey, Ph. D.

Date approved: May 6, 2015

Abstract

White students typically score higher on average than Black students on reading and mathematics tests and this gap appears to grow larger as students get older. This study used data from the national Early Childhood Longitudinal Study (ECLS) to follow about 7,400 students from 1st to 3rd to 5th grade and examined their performance on ECLS' standardized tests in reading and mathematics. Two sets of research questions were explored: 1) does the difference in mean performance increase across time?, and 2) does the difference in mean performance decrease when one controls for family, school and classroom variables? Analyses were conducted using a mixed analysis of variance and structural equation modeling techniques. Results found that the performance gap increased for reading from 1st to 3rd grade, but not from 3rd to 5th grade. For mathematics, the gap increased continuously from 1st to 3rd to 5th grades. The difference between mean scores in both reading and mathematics dropped substantially when family, school and classroom variables were taken into account.

Acknowledgments

I would never have been able to complete my dissertation without the guidance of my committee members and support from my family.

I would like to express my deepest gratitude to my advisor, Dr. Bruce Frey, for his excellent guidance, caring, patience, and providing me with the ultimate motivation throughout the entire period of my research. I would like to thank Dr. Vicki Peyton, who offered constructive suggestions and comments on the methodology and results of my analyses. I would also like to thank my minor advisor, Dr. Michael Imber, for helping me to develop my background in educational law and policy. I would also like to thank Dr. David Hansen, who honored my request without hesitation to join my dissertation committee after the withdrawal of Dr. Todd Little of the psychology department. I would like to thank Dr. Meagan Patterson for accepting to replace Dr. William Skorupski at the last minute.

I would like to thank Jill McCarroll, Education Statistician from the National Center for Education Statistics (NCES), for offering me excellent advice and for her recommendation on the appropriate weights to use for my analyses. I would also like to thank Terrence Jorgensen Jr. of Center for Research Methods and Data Analysis (CRMA) for offering advice on the limitation of NCES data used in the study.

I would also like to thank my daughter, Ariel Anib who continuously trusted, encouraged and motivated me to finish my dissertation. I would like to thank my six-year old son, Ray Niboro Jr. for his understanding when I asked him to read his book or play his games while working on the dissertation, even when he really needed my attention. I would also like to thank my son in Nigeria, Oluwaseun Aniboro for believing in me. Finally, I would like to thank my wife for her understanding and for being there for me even when I was unlovable. Thank you all!

Dedication

I wholeheartedly dedicate my dissertation to my mother, Mary Okakpokpo Niboro, who set my feet on the path of knowledge. She tenaciously encouraged me on the value of education as she continuously reminded me that, “Education is Power” and with education the sky is the limit. Unfortunately, she is not here to physically witness my graduation. I love you mom! May you Rest in Perfect Peace.

Table of Contents

Table of Figures

List of Tables

1 Introduction

Background of the problem

Addressing the difference in achievement between minority and non-minority students has become a national priority. In an era of No Child Left Behind (NCLB) and the Annual Yearly Progress (AYP), schools and districts not only are required to have all students meet minimum state standards, but they also need to make efforts to eliminate differentials in achievement among various groups (Behind, 2002; Simpson, Lacava, & Graner, 2004). For decades, achievement gaps have existed between Black and White students, and all efforts to eliminate this anomaly have proved elusive. In 1999, National Assessment of Educational Progress (NAEP) reports showed that White students had higher average test scores in reading and mathematics than their Black and Hispanic peers (Campbell, Hombo, & Mazzeo, 2000). The reports indicate that the gap between White and Black students in reading has narrowed between 1971 and 1999 in each age group; and has somewhat widened since 1988 at ages 13 and 17. Similarly, in mathematics the apparent gap has narrowed between 1973 and 1999 and has widened since 1986 at age 13.

When children enter early schooling, they perform differently upon evaluation for school readiness due to family economic differentials (Greg J Duncan & Katherine A Magnuson, 2005; Magnuson & Waldfogel, 2005). However, as learning occurs within the school context, ceteris paribus, there is a high likelihood that there would be a corresponding increase in cognitive mapping for all children, regardless of racial group. However, as students advance through grade levels, their academic performance is different based on their racial affiliation; White students outperform their Black counterparts, thereby creating an inequality in the educational outcome popularly known as the “achievement gap." This dissertation will address specifically Black-

White test score gaps in reading and mathematics and evaluate the extent to which family, classroom and school covariates affect the gap during the first five years of schooling.

Statement of the problem

The study will investigate test score differentials among Black and White students and the role of race on academic performance across three waves of data collection. The analysis will further explore whether family, classroom and school characteristics account for student performance and the extent to which these factors contribute to the apparent Black-White test score differentials.

Research hypotheses

A test score differential commonly known as an achievement gap is a phenomenon where the majority group outperforms the minority group on standardized tests in spite of exposure to congruent learning conditions and resources. Test score differential as it relates to Black and White might be different from Hispanic and White or Asian and White students; however, the premise is the same; the majority group outperforms the minority group. An array of research studies, for example Fryer and Levitt (2006) documented that Black students underperform academically when compared with their White counterparts. They posited that White students on the average score 0.274 standard deviation above the mean on the mathematics exam in fall kindergarten, whereas Black students perform 0.364 below the mean on that test, yielding Black-White gap of 0.638 standard deviation.by spring of first grade. Another study posited cognitive inequality between Black and White as a reason for the apparent gap, whereas others focused on inequality of environmental conditions as reasons for the gap. Herrnstein and Murray's (1995) study claimed that there is an association between family background and young children's cognitive skills. They argued that family characteristics for Blacks would have to be at the 6th percentile of the distribution of Whites in order for the test gap to be completely environmental.

The primary purpose of this study is two-fold: 1) to explore test sore differentials among Black and White students, and 2) to explicate the underpinning factors that contribute to Black-White achievement gaps. This study underscores differential outcomes in mathematics and reading where White students significantly outperform their Black counterparts from first grade through fifth grade. Specifically, this study addresses the following hypotheses:

1. The magnitude of the Black-White test score differential in mathematics increases from first through fifth grade.
2. The magnitude of the Black-White test score differential in reading increases from first through fifth grade.
3. Black-White performance gap in mathematics can be accounted for by family, classroom, and school variables.
4. Black-White performance gap in reading can be accounted for by family, classroom, and school variables.

Addressing these research hypotheses will provide researchers and policymakers with a possible explanation on the extent of the gap and the effects covariates have on the apparent gap.

2 Review of Literature

Historical Perspectives and Previous Literature

The research hypotheses addressed in this study focused on relationship of family, classroom, and school characteristics on race and Black-White test score gaps. The study illuminates the role race play on these covariates and the apparent gaps. However, before examining the effect of each factor on the gap, the study will expound historical perspectives and previous literature that address different outcomes among different racial groups.

The Coleman report (Coleman et al., 1966) was a national study that documented racial differences in academic achievement among children at various levels of schooling. It posited that achievement gap not only existed between Blacks and Whites at every grade level, but increased with student age (R. G. Fryer & Levitt, 2004). Additional research findings asserted that these disparities have been observed to exist before children enter kindergarten, widen as they move through middle schools, and persist into adulthood (Phillips, Crouse, & Ralph, 1998). Since then, several policies and reform agendas have been initiated by politicians to address the issue of divergence of scores among Black and White students.

The most current of these reforms is the NCLB that was inaugurated to hold districts and schools accountable to the same state academic achievement standards for all students (Simpson et al., 2004), regardless of racial or ethnic membership. The purpose of Title I clause of the NCLB (2002) is to ensure that all children have fair and equal opportunity to receive higher education; accomplished in part by closing the achievement gaps between minority and non-minority students. Since the implementation of the NCLB Act of 2001, closing achievement gaps between advantaged and disadvantaged students and non-minority and minority students has not only become a national priority, but it has remained a driving force that shapes the operation of educational policies and reform agendas across the country. An array of research endeavors have focused attention on understanding divergent trajectories of test scores between Blacks and Whites and how these gaps have changed over time, as well as the role of race, family, classroom, and school on the achievement gap. Each of these covariates and their effects on the Black-White gap will be discussed below.

Effect of family on Black-White gap

It is impossible to dissociate completely the effect of family on Black-White test score gap from the generality of student outcomes without consideration to the direct or indirect interaction of family context with either the classroom or the school contexts. This section focuses attention exclusively on family characteristics that have a profound effect on Black-White achievement gap. A number of facts have emerged. A high correlation exists between socioeconomic status, determined by parent education, income, and occupation, and family structure with the child's achievement.

Socioeconomic status and the effects of poverty are important factors in explaining racial differences in educational achievement (J. Brooks-Gunn, Duncan, & Klebanov, 1994, 1995; J. Brooks-Gunn & Duncan, 1997; J. Brooks-Gunn & Duncan, 2000). Socioeconomic status indicators are important metrics for evaluating cognitive preparedness of the child upon entering kindergarten and therefore are critical in the Black-White achievement gap. Several research studies have posited that Black students are disadvantaged with relative to financial resources, which has consequential effects on the Black-White gap. Duncan and Magnuson (2005) assert that the average socioeconomic level of Black kindergartners was more than two-thirds of a standard deviation below that of Whites. As a result of low socioeconomic status of Black students, their parents are unable to afford additional resources that would enhance their cognitive mapping, and consequently enable them to compete on the same level relative to White students.

Family environmental indicators, including mother's education, mother's income, household size, mother's perceived self-efficacy, and mother's parenting practices, are important ingredients in the Black-White achievement gap debate. Although, parent education and income are highly correlated, each indicator affects the child in different ways. According to Duncan and Magnuson (2005), higher family incomes might give a big edge in academic achievement. The study posits that higher income allows parents to provide their children with a stimulating environment, such as providing books, newspapers, and computers in the home. Several research studies documented that Blacks are much more disadvantaged on these family indicators than their white counterparts. Jencks and Phillips (Jencks & Phillips, 1998a) find that racial inequalities on these measures account for more than half the test score gap between Black and White five- and six-year-olds.

As previously discussed, parents' education is highly correlated with children's cognitive development and parenting styles (G. J. Duncan & K. A. Magnuson, 2005). Children with highly educated parents roughly score higher on cognitive and academic achievement tests than do children of parents with less education. Duncan and Magnuson (2005) point out that the link between children's cognitive development and parental education is evident as early as in the child's life as three months of age.

Many researchers have also argued that the major factor that explains the differences in student achievement has to do with disparities in material resources and conditions that exist among students, their families, and their schools (Armor, 2003; Rothstein, 2004). For example, language minority students who are not proficient in English begin school with lower levels of achievement and progress more slowly in school than students from English-only backgrounds (Gandara, Rumberger, Maxwell-Jolly, & Callahan, 2003). Black students are known to have limited amount of vocabulary prior to entering kindergarten that prevent them from competing academically on a level commensurate with their White counterparts. Refugees' children especially those who made an exodus from Africa to United States, have a limited English proficiency that further exacerbates the Black-White gap.

Another important indicator of family background is family structure. Today about one-third of all children are born outside marriage, and more than half of all children will live in single-parent family at some point in their childhood (G. J. Duncan & K. A. Magnuson, 2005). Duncan and Magnuson (2005) assert that young children living with single mothers face poverty at five times the rate of preschoolers in intact families (50% versus 10%), and the declines in income for households with children after a divorce are dramatic and lasting. Research evidence showed that children of single-parent families face worse financial deprivation than children of two-parent families, and account for a substantial proportion of children's achievement.

Effects of classroom and teacher on Black-White gap

This section addresses classroom- and teacher-related characteristics that have profound effects on the Black-White test score differentials. The first characteristics address the effects of teacher's perceptions and expectations on the gap, while the second address research that posit the effects of teacher's qualifications and experiences on the Black-White test score divergence.

Teacher's perceptions and expectations

Ferguson (2003) examines evidence for the proposition that teachers' perceptions, expectations, and behaviors interact with student' beliefs, behaviors, and work habits in ways that help perpetuate the Black-White test score gap. Ideally, upon entering kindergarten, the classroom and school processes interact to shape the child's cognitive mapping and equalize achievement across ethnic groups. Unfortunately, national data have shown that the Black-White gap either remains constant (in standard deviation units) or widens as the child transitions from elementary to secondary grades (Fryer Jr & Levitt, 2004).

Teachers' perceptions and expectations toward individual students based on their race or ethnicity may be biased or unbiased. Bias is a deviation from some benchmark that defines neutrality, or lack of bias (Phillips et al., 1998). Inconsistencies have been found in the results of research studies on this subject. A meta-analysis of experimental studies found that in nine studies, teachers have higher expectations for White students, and they have higher expectations for Blacks in one study (Baron, Tom, & Cooper, 1985). However, other studies do not report which group is favored.

Teachers' qualifications

The NCLB legislation of 2001, in part requires that schools and districts recruit highly qualified teachers with at least a bachelor's degree in their classrooms. A number of recent studies have shown that teacher qualifications have a significantly positive relationship with academic achievement. Using data across grade levels aggregated at the state level, Darling-Hammond (2000) found that before and after controlling for poverty and language status, teacher certification had a stronger correlation with reading achievement than did class size, teacher salaries, or school spending. Several other research studies (Ferguson, 1998; Goldhaber, 2002; Rivkin, Hanushek, & Kain, 1998; Sanders, 1998) found that teacher characteristics are more predictive of student achievement than are school characteristics.

Using data from Texas, Hanushek, Kain, and Rivkin (2002) found that teacher differences accounted for a minimum of 4% of the variance in achievement. Also, using data from Tennessee, Sanders and Rivers (1996) found that the norm referenced test scores of students with most effective teachers increased 36 percentile points more than students with the least effective teachers. Although there is sufficient evidence to support effects of teacher characteristics on student achievement, there was no evidence to show that either teacher qualifications or experience has any effect on Black-White test-score gap.

Effects of school on Black-White gap

Coleman et al. (1966) and Borman and Dowling (2010) found that both school racial composition and social class composition influence student educational outcomes. The difference in educational outcomes between Black and White students and how to bridge the apparent gap continue to be a challenge in the educational and research communities. Academic success is determined by the interactions between family, classroom (or teacher) and overall schools factors. Studies on the effect of schools on Black-White gap reveal two important views. Black children fall behind their White counterparts in reading and vocabulary before they enter kindergarten, and they learn less than White students through K-8 schooling (Jencks & Phillips, 1998a).

Because race is highly correlated with socioeconomic status, the effect of school racial composition could be the result of concentrated poverty (Kahlenberg, 2001; Rumberger & Palardy, 2005). That is, most of the variation in student achievement is attributable to differences in between students (and their families), rather than differences between schools (Lee & Bryk, 1989; Reardon, 2003; Rumberger & Palardy, 2004). Even when Black and White children enter school with the same skills, Black children learn less new skills than Whites, making Blacks fall further behind academically.

White students produce significantly higher achievement scores than Black students at all levels of education (Jencks & Phillips, 1998d; NCES, 2000, 2001). Bali and Alvarez (2004) documented that the Black-White achievement gap begins before the first grade. This gap continues to grow as student progress through the school system (Bali & Alvarez, 2004; Jencks

& Phillips, 1998a). By the time minority students reach high school, the gap has grown so much that Black student achievement may be as much as 0.34 standard deviations below population mean (Phillips et al., 1998).

The dissertation explores the relationship among family, classroom, and school characteristics and race on test score gap. Black-White achievement gaps continue to exist despite the landmark case of Brown vs. Board of Education (1954), an equality initiative that focused on educational opportunities for children of color. The intent of the initiative was that all children, whether Black or White, will receive equal educational opportunities. For over six decades since Brown's decision, studies have focused on differentials in achievement scores among Black and White students. A reasonable array of literature documented that the Black youths have lower academic achievement (e.g., standardized test scores) than White students and that this difference exists across various stages of schooling (Downey & Pribesh, 2004; Downey, von Hippel, & Broh, 2004; Reardon, 2003; Roscigno, 1998; Scarr, 1981).

This study examines the nature of Black-White reading and mathematics achievement gaps, including the extent of the differences to which these gaps are associated with individual, family background, classroom, and school characteristics. The study will explicate differential outcomes in reading and mathematics among Black and White students from first through fifth grade. Specifically, this study addresses the following underlining hypotheses:

1. The magnitude of the Black-White test score differential in mathematics increases from first through fifth grade.
2. The magnitude of the Black-White test score differential in reading increases from first through fifth grade.

3. Black-White performance gap in mathematics can be accounted for by family, classroom, and school variables.
4. Black-White performance gap in reading can be accounted for by family, classroom, and school variables.

3 Methods

Participants

The Early Childhood Longitudinal Survey Kindergarten (ECLS-K) is a nationally representative sample of 21,260 children enrolled in 944 Kindergarten programs during the 1998-1999 school year that was designed to study the development of educational stratification among Unites States school children (West, Denton, & Germino-Hausken, 2000). The study was developed by the U.S. Department of Education (USDE), National Center for Education Statistics (NCES) and implemented by Westat, a research corporation based in Rockville, Maryland with assistance provided by Educational Testing Service (ETS) in Princeton, New Jersey. NCES recommended removing five students from the sample for substantial data errors (Tourangeau et al., 2009), thus leaving a sample of 21,255 children for the analyses in this study.

Sample selection for the ECLS-K involved a dual-frame, multistage sampling design (Tourangeau et al., 2009). At the first stage, 100 primary sampling units (PSUs) were selected from a national sample of PSUs that comprised of counties and county groups. At the second stage, public schools were selected within the PSUs from the NCES Common Core of Data (CCD), “Public Elementary/Secondary School Universe Survey,” and the NCES Private School Universe Survey (PSS). An additional frame called “freshened” was created to identify kindergartners that were included in the original frame. Altogether, 1,280 schools were sampled from the original frame and 133 from the freshened frame. Of these, 953 were public schools and 460 were private schools (Tourangeau et al., 2009). Children of Asian and Pacific Islanders (APIs) were oversampled. Schools participated with a weighted response rate of 74%; among the participating schools, the completion rates were 92% for the children, 91% for the teachers, and 89% for the parents (Raver, Aber, & Gershoff, 2007).

Prior to beginning this dissertation and to protect the rights, well-being and privacy of individual subjects in the data structure, and to ensure the interests of the University of Kansas and in compliance with the Public Health Service Act (Pub. L. 93-348) as amended, the study seek approval from the Human Subjects Committee - Lawrence (HSCL). The primary mission of HSCL is to protect participants' rights and privacy; and also protects researchers from legal and ethical mistakes and safeguards them from the consequences of such mistakes.

Definition of variables

Two types of variables, a dependent variable and several independent variables were explored to better understand the process and complexity of the study. Dependent variable means the response that is being measured. In this study, direct cognitive assessment scores for reading and mathematics were used as a dependent variable.

Dependent variables

The reading and mathematics achievement scores used in this study are item response theory (IRT) scale scores extracted from the ECLS-K 1998-99 data collection for three time points; grades 1, 2 and 3. The data used in this dissertation were collected in spring first grade (1999-2000); spring third grade (2001-02), and spring fifth grade (2003-04). The scores have the ability to illustrate increase, decrease, or no change in achievement gaps among Black and White subgroups over time.

IRT uses the pattern of right, wrong, and omitted responses to the items actually administered in an assessment and the difficulty, discriminating ability, and "guess ability" of each item to place each child on a continuous ability scale. IRT procedures use the pattern of responses to estimate the probability of correct responses for all assessment items; and scoring makes possible longitudinal measurement of gain in achievement over time, even though the assessments that are administered are not identical at each time point (Tourangeau et al., 2009).

Additionally, scoring will also make it possible to establish divergence in achievement between Black and White students during the first five years of schooling. Reading and mathematics assessments were designed to measure specific skills.

Independent variables

The second categories of variables are the independent variables. Independent variables are variables that are varied or manipulated by the researcher. Independent variables used in this study are race/ethnicity (Black and White), family background, classroom (or teacher), and school characteristics. Family background characteristics include but not limited to socioeconomic status based on father's education, mother's education, father's occupation, mother's occupation, and household income. Classroom characteristics include qualifications of individual teachers, years of experience, classroom climate and environment (parceled); while school characteristics are school enrolment, percent of reduced lunch and free lunch eligible students.

Family characteristics

Family characteristics encompass race of child, family type, number of siblings, socioeconomic status (SES) and poverty, parent education, father’s education, mother’s education, father’s occupational prestige, mother’s occupational prestige, and home educational resources. Home resources include number of books at home, computer at home, visits to the library, zoo or aquarium, and museum. These resources enhance cognitive skills of the child prior to beginning kindergarten. Race/ethnicity information on child was collected from the parent interview data. Race indicates that the child belonged to one or more of the following race and ethnic categories of “Hispanic,” “White,” “Black,” “Asian,” and “other” (Zill & West, 2001). Altogether, there were 21409 participants in this study. Of these, 11788 (or 55.1%) were White, 3224 (15.1%) were Black, 3826 (17.9%) were Hispanic, 1366 (6.4%) were Asian, and

1154 (5.4%) were categorized as other. To provide the race/ethnicity of the child, parents completed a set of questions regarding their classification of their child racial and ethnic group. In this study, data points were analyzed based on two racial groups, Black and White, and all other ethnic groups were excluded from the analyses.

The SES of participants was computed at the household level based on data collected from parents' interview in eighth grade (Fall 2006). The components used to create socioeconomic profile of the child were the father's education, mother's education, father's occupation, mother's occupation, and household income (Tourangeau et al., 2009). Home educational resources represent physical home environment and cognitively stimulating materials (Yeung & Conley, 2008), such as educational toys, books, and home computers. In this study, home resources are defined as those resources provided by parents at home to give additional cognitive enhancements to the child. These resources may also include trips to the zoos, museums, and libraries in order to provide additional cognitive enhancements. Table 1 shows family variables and estimated means and standard deviations for full sample and Black and White groups for each variable.

Table 1: Summary statistics of family characteristics of full sample and by race

Variable	Full sample	White	Black
Race	.21(.41)	.00(.00)	1.00(.00)
Family type	1.64(1.03)	1.51(.93)	2.44(1.28)
Poverty level	1.89(.32)	1.93(.26)	1.62(.49)
Household income	.10(.77)	.19(.74)	-.47(.73)
Father's occupational prestige score	43.34(14.72)	43.96(14.47)	36.16(15.68)
Mother's occupational prestige score	38.40(20.11)	38.93(20.31)	34.94(18.38)
Father's highest education level	5.48(1.85)	5.65(1.82)	4.39(1.62)
Mother's education level	5.12(1.95)	5.19(1.95)	4.25((1.71)
Parent's education level	5.02(1.72)	5.14(1.70)	4.21(1.57)
Number of children's books at home	120.19(182.78)	130.76(189.82)	61.49(121.56)
Have home computer child uses	1.11(.31)	1.07(.26)	1.32(.47)
Visited zoo or aquarium	1.70(.46)	1.71(.46)	1.68(.47)
Visited a museum	1.66(.48)	1.64(.48)	1.72(.45)
Visited library	1.52(.50)	1.52(.50)	1.51(.50)

Notes: The entries are means and standard deviations of family-level data for students in ECLS-K who do not have missing values for race. Race - dummy coded as 0 for white and 1 for black. The category white includes only non-Hispanic whites. Family type categories using both parent and sibling information. Family income is classified as $25,000 and less per year, or as greater than $25,000. Table 2 shows ECLS-K's household income compared to census poverty thresholds for 2006. A household that has income below the appropriate threshold is considered poor. Table 3 shows father's and mother's occupation General Social Survey (GSS) prestige score computed as the average of corresponding prestige values for the 1980 occupational categories. Parent's/mother's/father's highest level of education from 8th grade or below to doctorate or professional degree. The number of books child has at home ranging from 0 to 3,000 books. The categories of home computer child uses, visitation to a zoo, aquarium, museum, or library coded as 1 for yes and 2 for no.

Table 2: ECLS-K and census thresholds for 2006: School year 2006-07

Household size	ECLS-K income categories	Census weighted average threshold for 2006
2	Less than or equal to$15,000	$13,167
3	Less than or equal to$20,000	$16,079
4	Less than or equal to$25,000	$20,614
5	Less than or equal to$30,000	$24,382
6	Less than or equal to$35,000	$27,560
7	Less than or equal to$40,000	$31,205
8	Less than or equal to$45,000	$34,774
9	Less than or equal to$50,000	$41,499

Note: U.S. Census Bureau, Current Population Survey.
http://www.census.gov/hhes/www/poverty/threshold/thresh06.html
Source: U.S. Department of Education, National Center for Educational Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

Table 3: Father's and mother's occupation prestige derived from 1989 GSS prestige score

Variable	Occupation GSS prestige score
Father’s/mother’s occupation GSS prestige score	29.6 Handler, Equip, Cleaner, Helpers, Labor
	33.42 Production Working Occupation
	34.95 Service Occupations
	35.63 Agriculture, Forestry, fishing Occupations
	35.78 Marketing & Sales Occupation
	35.92 Transportation, Material Moving
	37.67 Precision Production Occupation
	38.18 Administrative Support, Including Clark
	39.18 Mechanics & Repairs
	39.2 Constructive & extractive Occupations
	48.69 Technologists, Except Health
	52.54 Writers, Artists, Entertainers, Athletes
	53.5 Executive, Admin, Managerial Occupation
	57.83 Health Technologists & Technicians
	59 Social Scientist/Workers, Lawyers
	61.56 Registered Nurses, Pharmacists
	62.87 Natural Scientists & Mathematicians
	64.43Teacher, Except Postsecondary
	64.89 Engineers, Surveyors & Architects
	72.1 Teachers, College Postsecondary Counselors, Librarians
	77.5 Physicians, Dentists, Veterinarians

Source: U.S. Department of Education, national Center for Educational Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), spring 2007.

Teacher and classroom characteristics

Table 4 shows summary statistics of teacher and classroom characteristics. It includes teacher's highest level of education, years of experience, class enrollment, and number of computers in class. Additional variables parceled as classroom's climate and environment are classroom's space and size, lighting, ventilation, condition, temperature, and noise level.

Table 4: Summary statistics of teacher and classroom characteristics of full sample and by race

Variable	Full sample	White	Black
Number of years been school teacher	15.35(10.18)	15.63(10.10)	13.92(10.46)
Highest education level teacher achieved	2.20(.94)	2.23(.92)	2.07(1.01)
Number of students in class	21.00(4.18)	21.10(4.18)	20.50(4.12)
Number of computers in class	2.83(1.89)	2.83(1.90)	2.82(1.86)
Classroom climate and environment	1.03(.12)	1.03(.11)	1.05(.14)

Note. The entries are means and standard deviations of teacher/classroom-level data for students in ECLS-K who do not have missing values for race. The number of been a school teacher variable ranges from one to thirty five years. Highest education level achieved by teacher is coded 1 for bachelor's degree through 4 for doctorate. The number of computers available in teacher's classroom ranges from zero to a maximum of eight computers. Finally, classroom's climate and environment is a parceled variable that encompasses classroom's size and space, lighting, temperature, condition, ventilation, and noise level. This variable is codded 1 = favorable and 0 = unfavorable for a classroom's climate and environment.

School characteristics

Table 5 shows summary statistics of school characteristics. It includes total school enrollment, percent of free lunch eligible students, and percent of reduced lunch eligible students.

Table 5: Summary statistics of school characteristics of full sample and by race

Variable	Full sample	White	Black
Total school enrollment	3.71(1.17)	3.69(1.18)	3.86(1.11)
Percent of free lunch eligible students	26.55(25.71)	20.44(19.90)	56.05(29.84)
Percent of reduced lunch eligible students	2.76(1.06)	2.71(1.06)	3.00(1.11)

Notes: The total school enrolment (TSE) is coded with “1” representing 0 -149 students through “5” representing 750 or more. Summary statistics indicate that TSE, White and Black students’ enrollment range from 300 to 499. The percent of free lunch eligible students range from a minimum of 0% to a maximum of 95%, where “0%” indicates full-paid lunch. Approximately 20% of White students are eligible for free lunch compared to 56% of their Black counterparts. Percent of reduced lunch eligible students is coded as “1” for less than 1% to “5” for 25% or more. Summary statistics indicate that about 5 to 10% of students, White or Black are eligible for reduced lunch.

Instrumentation

This section addresses primary instruments used in this study to assess children’s academic development. It comprises of the cognitive assessment instruments, reliability of the instruments and validity.

Cognitive assessment instruments

The ECLS-K (1998-99) assessment instruments were designed to assess children’s academic and social development during the kindergarten through eighth grade years (Tourangeau et al., 2009). The study used direct cognitive measures that describe children’s academic performance at each time point, as well as growth trajectory over time. Direct cognitive assessment battery was designed to assess children’s academic achievement in the spring of eighth grade , 2006-2007 school year, and to provide a means of measuring academic growth in three subject domains, reading, mathematics, and science, that had been part of the child’s educational experience since kindergarten entry (Tourangeau et al., 2009).

The ECLS-K total item pool consisted of 212 items for reading, 174 items for mathematics, and 111 items for science. Of these, 10 to 25 were taken by all children within each round of data collection (Tourangeau et al., 2009). In order to measure growth across time, cognitive assessments were designed to have overlapping items, that is, items were retained across time points to support the development of longitudinal score scales in each subject area (Najarian et al., 2009; Tourangeau et al., 2009). For the purpose of this study, only three data points for mathematics and reading grades 1, 3 and 5 were used.

Development of the K-1-3-5-8 longitudinal scale

Scale scores were linked across grade and require both overlapping ability distributions and overlapping test forms (Najarian et al., 2009). The reliance on common items that were present in more than one set of test forms allowed the development of a vertical scale appropriate for measuring the divergence of mathematics and reading scores between Black and White students. Examinee performance on the items that are common to adjacent test levels are used to indicate the amount of growth that occurs from one grade to the next (Brennan, 2004). In vertical scaling, tests that differ in difficulty, but that are intended to measure similar constructs are placed on the same metric.

Reliability

The reliability of an instrument is the consistency of measurement under different circumstances. That is, it is the degree to which individuals' deviation scores, or z-scores, remains relatively consistent over repeated administration of the same test or alternate test forms (Algina & Crocker, 2006). Cronbach (1951) presented a synthesis and discussion on various methods of estimating internal consistency represented in a general formula known as Cronbach’s alpha. Cronbach’s alpha, also known as coefficient alpha (α) can be represented mathematically as

$$\alpha \equiv \frac{n}{n-1}\left[1 - \frac{\sum \sigma 2(Yi)}{\sigma 2x}\right] \tag{1}$$

where n is the number of items on the test, σ^2 is the variance of item *i*, Y_i is the measurement of item *i*, and σ^2x is the total test variance.

In general, the more items a test has and the greater the ability of test takers, the greater the reliability is likely to be. The most appropriate estimate of the reliability of the assessment is the reliability of the overall item response theory (IRT) ability estimate, theta, since it reflects the internal consistency of performance. Table 6 shows the reliability of IRT-based scores by round of data collection and domain (Tourangeau et al., 2009) for grades 1, 3 and 5.

Table 6: Reading and mathematics assessment reliabilities, spring first grade through spring fifth grade: School years 1999-2000, 2001-02, and 2003-04

Domain	Spring first grade	Spring third grade	Spring fifth grade
Reading	.96	.94	.93
Mathematics	.94	.94	.95

Notes: The table presents unweighted reliability statistics of theta. These are reliabilities of theta scores based on IRT-scale scores - number of correct or incorrect answers to obtain estimates on a vertical scaling that may be compared in different assessment forms.
Source: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998 - spring 2007 (Adapted).
.

Relative to the IRT-based scores, the reliability of the overall ability estimate, theta, is based on the variance of repeated estimates of theta compared with total sample variance (Tourangeau et al., 2009). For example in the spring first grade, a reading of $\alpha = .96$ reveals that at least 96% of the total score variance is due to true score variance, and mathematics of $\alpha = .94$ indicates that at least 94% of the total variance is attributed to true score variance.

Validity

Cronbach (1971) described validity as the process by which a test developer or test user collects evidence to support the types of inferences that are to be drawn from test scores (Algina & Crocker, 2006). The ECLS-K evidence for the validity used in this study were derived from different sources. They were derived from national and state performance standards, comparison with state and commercial assessments, and judgments of curriculum experts who all provided input to test specifications (Tourangeau et al., 2009). Assessment items were drawn from assessments used in other large-scale studies, such as National Assessment Educational Progress (NAEP), the National Education Longitudinal Study of 1988 (NELS: 88), and Educational Longitudinal Study of 2002 (ELS: 2002).

The dissertation addressed two types of validity: content and criterion validity. Content validity is where the test user desires to draw an inference from the examinee's test score to a larger domain of items similar to those on the test itself. While criterion validity is where the test user desires to draw an inference from the examinee's test score to performance on some real behavioral variable of practical importance. For this study, assessment data for mathematics and reading were extracted for three time points, spring first, third, and fifth grades. These multiple rounds of data collection can give valuable information about the trends in the child's academic development from first through fifth grades.

Data cleaning and analyses

SPSS Statistics Version 21 software package (IBM, 2010) was used to extract publicly available data from NCES.ed.gov and for initial data cleaning. Before proceeding with the analysis, all variables were screened for possible code and statistical assumption violations, as well as for missing values and outliers, with SPSS Frequencies, Explore, Plot, Missing Value Analysis, and Regression procedures. First, three separate data sets were created, one for Black, another for White, and lastly a combination of Black and White racial groups. In each data set, missing data points and invalid data responses were transformed and coded as 999. Lastly, the newly created data sets were saved in a comma delimited format with a file extension "csv" to be used by M*plus* Version 6.1 statistical package (Muthén & Muthén, 2010) for evaluating hypotheses 3 and 4. Before continuing with the data analysis all variables were screened as follow: addressing missing values, dealing with univariate outliers, assessing assumptions, and checking for multivariate outliers.

Missing data

There are two main reasons for missingness. One of the main methodological problems in longitudinal studies is attrition (Twisk & de Vente, 2002). Attrition, missing data, or dropouts; all terms are used for the situation that not all N subjects have data on all repeated measurements. The second reason for missingness is participants' nonresponse to survey items. ECLS-K data set was associated with extensive amount of missingness as a result of nonresponse (Schlomer, Bauman, & Card, 2010) and attrition due to relocation of study participants to another country or simply refused to complete the study. As ECLS-K data collection progressed from first wave to the seventh, the amount of missingness increased astronomically, from 6.6% at time point 1 to 56.6% at time point 7. To account for missingness in this study, all variables in ECLS-K data used a standard scheme for missing values (Tourangeau et al., 2009). Table 7 shows codes that were used to indicate item nonresponse, legitimate skips, and unit nonresponse. For the purpose of this analysis, all codes (i.e. -1, -7, -8, -9, and blanks) were translated into system missing or periods (".") and later recoded into "999" to be used in M*plus* analysis.

Table 7: Missing values codes, School years 1999-2000, 2001-02 and 2003-04

Value	Description
-1	Not applicable, including legitimate skips
-7	Refused (a type of item nonresponse)
-8	Don't know (a type of item nonresponse)
-9	Not ascertained (a type of item nonresponse)
Blank	System missing, including unit nonresponse

Source: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), fall 1998 through spring 2007.

All the 15012 participants were screened for missing values on twenty three continuous variables (reading grades 1, 3, and 5; mathematics grades 1, 3 and 5; NUMSIB, TYPFAMIL, MOMED, DADED, PARED, MOMSCR, DADSCR, SES, POVRTY, CHLBOO, CLSIZE, COMPUT, TREXP, TRED, SCHENRLS, FLNCH, and RLNCH) and six categorical variables (RACE, LIBRAR, MUSEUM, ZOO, HOMECM, and CLENV). All variables were screened using the full sample and subpopulations of White and Black students.

No missing data was detected for race. Of the six waves of data collection for the dependent variable of reading and mathematics, only three time points each were used for ANOVA and MIMIC analyses. These were time points for grades 1, 3 and 5 that were approximately equidistant from each other. An inspection of the univariate statistics for reading shows that grade 1 variable had 3235 (21.5%) missing values with a mean of 78.68 (SD = 23.84), based on 11777 valid cases. Grade 3 variable had 5074 (33.8%) missing values with a mean of 130.12 (SD = 27.47), based on 9938 valid cases. Finally, Grade 5 variable had 7273 (48.4%) missing values with a mean of 153.43 (SD = 25.53), based on 7739 valid cases. The data matrix for mathematics shows that grade 1 variable had 3,236 (21.6%) missing values with a mean of 62.97 (SD = 18.32), based on 11,776 valid cases. Grades 3 and 5 variables had 5,015 (33.4%) and 7,267 (48.4%) missing values, respectively. Pairwise linearity was deemed satisfactory because scatterplots were oval shaped indicating linearity between two variables. All missing values were replaced based on multiple imputation algorithm using M*plus (*Muthén *&* Muthén, 2010*)*.

Pattern of missing data

There are three patterns of missingness: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR), also known as nonignorable nonresponse. With MCAR data, there are no patterns in the missing data and the missing values are not related to any variables under study (Acock, 2005; Bennett, 2001; Roth, 1994). In a particular data matrix, missing data points are randomly distributed throughout the matrix. That is, the missing data mechanism is characterized by the conditional distribution of M (missing data) given Y (observed data), $f(M|Y, \emptyset)$, where $\emptyset$ denotes unknown parameters (R. J. Little & Rubin, 2014). Mathematically,

$$f(M|Y,\emptyset) = f(M|\emptyset) \text{ for all } Y, \emptyset \tag{2}$$

The second type of missing data mechanism is MAR. With MAR data, the probability of having a missing data point is related to another variable in the data matrix but is not related to the variable of interest (Allison, 2001). Missing data are related to observed data (another variable in the data set) but not to missing data (Graham, Cumsille, & Elek-Fisk, 2003; Roth, 1994; Schafer & Graham, 2002). That is, missingness depends only on the components Y_{obs} of Y that are observed, and not on the components that are missing, Y_{miss} (R. J. Little & Rubin, 2014). Mathematically,

$$f(M|Y, \emptyset) = f(M|Y_{obs}, \emptyset) \text{ for all } Y_{miss}, \emptyset \tag{3}$$

The third type of missing data mechanism is NMAR. Here, the likelihood of missingness is related to the score on that same variable had the participant responded. That is, the distribution of M depends on the missing values in the data matrix, Y. Mathematically,

$$f(Y, M|\theta,\emptyset) = f(Y|\theta) f(M|Y, \emptyset) = \prod_{i=1}^{n} f(yi|\theta) \prod_{i-1}^{n} f(M|yi,\emptyset) \tag{4}$$

Missing data can become problematic for social and education researchers. Missingness is more severe in longitudinal studies with multiple time points where data is collected on the same participant multiple times. The data missing mechanism associated with ECLS-K data appears to be MAR for three reasons. First, a close observation of the data matrix revealed that there was a systematic relationship between one or more of the variables and the probability of missing data (Enders, 2010). Second, ECLS-K coding scheme revealed legitimate skips by respondents resulting to nonresponse or missingness as shown in Table 7. Finally, Little's MCAR test conducted using the missing value analysis (MVA) option of SPSS (IBM, 2010) showed $p < .05$, indicating a MAR missing mechanism. As a result of missingness, statistical analyses are likely to be biased. To minimize the problem of missinngness, a variety of strategies are used for handling missing data. These methods are grouped into deletion, nonstochastic imputation and stochastic imputation methods. This dissertation used one of the stochastic imputation methods known as multiple imputation (MI).

MI refers to the procedure of replacing each missing value by a vector of $D \geq 2$ imputed values (R. J. Little & Rubin, 2014), where imputation values, M are calculated for every missing value (Twisk & de Vente, 2002). MI involves a three-step process: 1) create 5 or 10 data sets using augmentation, 2) estimate the model (e.g. regression, logistic regression, SEM) separately for each of the 5 or 10 data sets using augmentation, and 3) compute pooled estimates of the parameters and standard errors using the 5 or 10 solutions. This dissertation used M*plus* (Muthén & Muthén, 2010) to generate 10 data sets, estimated the SEM model separately for each of the data sets and computed pooled estimates of the parameters and standard errors using the 10 solutions. Appendix 1 shows M*plus* syntax for MI to create 10 data sets used for further analyses.

Altogether, fifty eight univariate outliers (17 for reading grade 1, 4 for reading grade 3, 12 for reading grade 5, 17 for mathematics grade 1, and 8 for mathematics grade 5) were detected, none of which were considered extreme or unusual enough to require deletion. The issue of univariate normality for the dependent variables was examined for skewness and kurtosis values. Table 8 shows that skewness and kurtosis values are within +1.0 and -1.0 range, therefore they were acceptable for the analysis.

Table 8: Descriptive statistics showing means, standard deviations, skewness, and kurtosis of dependent variables using full sample (N = 7466)

Variable	Mean/SD	Skewness/SES	Kurtosis/SEK
Reading grade 1	78.68(23.84)	.71(.02)	.48(.05)
Reading grade 3	130.12(27.47)	-.26(.03)	-.42(.05)
Reading grade 5	153.43(25.53)	-.59(.03)	.10(.06)
Mathematics grade 1	62.97(18.32)	.46(.02)	.32(.05)
Mathematics grade 3	100.92(24.45)	-.11(.02)	-.64(.05)
Mathematics grade 5	125.57(24.23)	-.70(.03)	-.00(.06)

Note: SD – Standard deviation, SES – Standard error of skewness, SEK – Standard error of kurtosis. SD, SES, and SEK are all in parentheses.

Amount of missing data

Experts have not reached a consensus regarding the percentage of missing data that becomes problematic (Schlomer et al., 2010). Schafer (1999) recommended 5% as the cutoff. However, Bennett (2001) suggested that when more than 10% of data is missing, statistical analyses are likely to be biased; and others have used 20% (Peng, Harwell, Liou, & Ehman, 2006). In contrast to those who suggested a particular cutoff, Schlomer et al. (2010) believe that two factors determine whether a certain amount of missingness is problematic. The first is whether the resultant data set has statistical power to detect the effects of interest, and second is the pattern of missingness. Simulations reported by Schafer (1997) show that with m = 10 imputations, when the MAR assumption is correct, MI is 94% as efficient as if there were no missing values when actually 50% of the values are missing. On close observation of the data matrix and the result of Little's MCAR test, this dissertation assumed a MAR data missing mechanism. Therefore with m = 10 imputations, resultant data gave sufficient power to detect Black-White test score differentials.

Data analytic strategy

The purposes of this dissertation were two-fold. The primary purpose was to investigate that test score differentials between Black and White students increase as they advance from first through fifth grade. The secondary purpose was to examine the effects of family, classroom (teacher) and school covariates on the Black-White gap. Analyses were conducted using a two-way repeated measures analysis of variance (ANOVA) and structural equation modeling (SEM) frameworks. A two by three repeated measures ANOVA was used to compare group means on a DV (IRT scale scores) across repeated measures of time (grades 1, 3, and 5). Time is often referred to as the within-subjects factor, whereas a fixed or non-changing variable (e.g. race) is referred to as between-subjects factor (Huck, Cormier, & Bounds, 1974). IRT scale scores having three measures (grades 1, 3 and 5) were the DVs; while race with two subgroups the IVs. In this study, differences in score variability (DV) were compared across time (within-subjects factor) by race (between-subjects factor). Two-way repeated measures ANOVA was used to test hypotheses 1 and 2 that the magnitude of test score differentials between White and Black students increased from first grade through fifth grade in reading and mathematics.

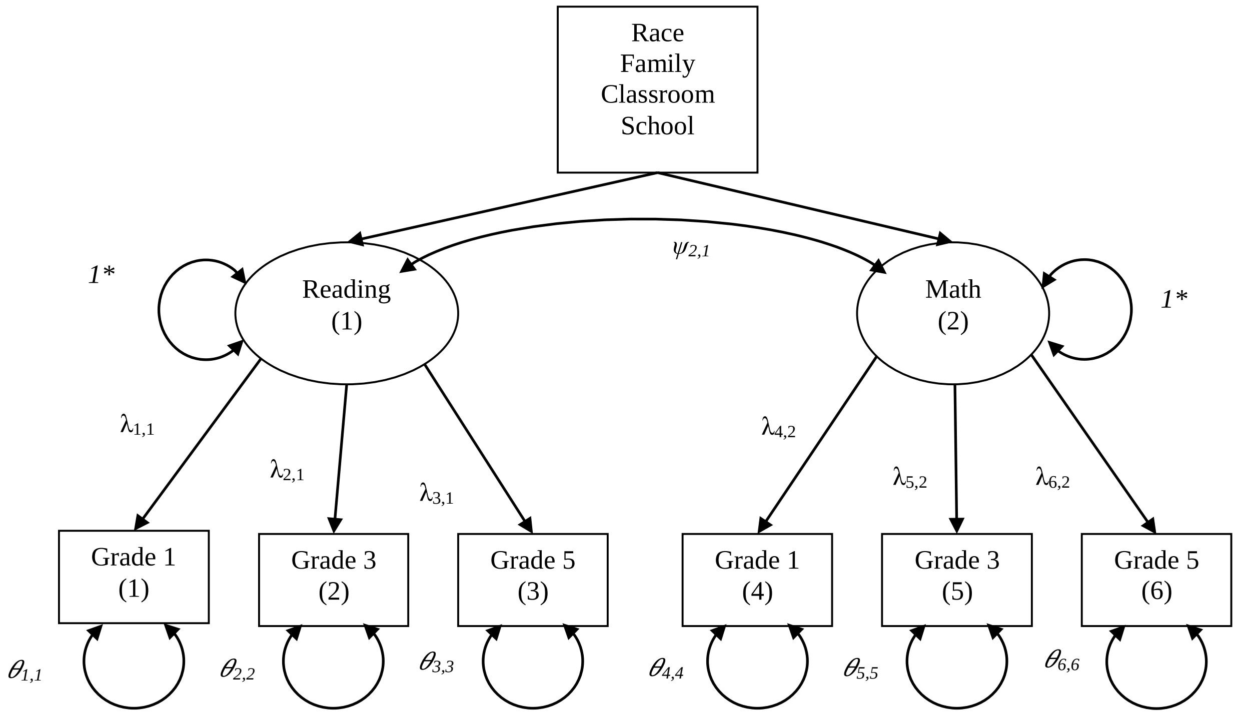
In the SEM framework, a confirmatory factor analysis (CFA) with covariates, popularly known as multiple indicators multiple causes (MIMIC) model (Jöreskog & Goldberger, 1975) was used to test hypotheses 3 and 4. This model entails regressing the latent factors and indicators onto covariates that represents the race membership. Unlike multiple-group CFA, a single input matrix was used in the analysis to evaluate the effect of family, classroom, and school characteristics on test score differentials between White and Black students. Specifically, MIMIC was used to evaluate if there was a decrease in test score differentials after controlling for family, classroom and school covariates in the model. MIMIC models were used to test whether the difference in performance in reading and mathematics between Black and White can be accounted for by family, classroom and school characteristics, assuming that the constructs were factorially invariant across groups.

In this study, reading and mathematics were used as the latent constructs, while the three time points (grades 1, 3 and 5) each for reading and mathematics with equal time intervals were used as the observed indicators. Because of the equal time intervals, it was appropriate to measure the growth trajectories from first through fifth grade. These time points were represented by IRT scale scores of student achievement scores in reading and mathematics for spring first, third, and fifth grades. Assessment items were administered in spring-first grade (1999-2000 school year), spring-third grade (2001-2002 school year), and spring-fifth grade (2003-2004 school year).

Investigating change in behavior across time (McArdle & Epstein, 1987; Meredith & Tisak, 1984, 1990) and interrelations across time among two or more processes (Cole & Maxwell, 2003; Curran & Bollen, 2001; S.E. Maxwell & Cole, 2007; McArdle, 2001; McArdle & Hamagami, 2001) are predicated on satisfying a key assumption: for each construct of interest, the study measured the same properties in the same metric at each occasion (Widaman et al., 2010). That is, the same trait or ability was measured by observed indicators across groups of race. MIMIC modeling was formally used to test two potential sources of factorial invariance. The first source of factorial invariance was by regressing indicators (grade 1, 3 and 5) onto covariates (family, classroom, and school factors) to test for the equivalence of indicator intercepts in Black and White groups. While the second was by regressing latent variables (reading and mathematics) onto covariates (family, classroom, and school factors) to test for the equivalence of factor means across groups.

Figure 1 shows a general path diagram used in each model. In each model presented in the study, the figure used standard figural notations for path diagrams: a) rectangles represent manifest variables; b) circles represent latent variables; c) one-headed arrows depict unidirectional direct effects of one variable on another; and d) double-headed arrows designate nondirectional relations between variables reflecting variances and covariances. For example, a double-headed arrow from a variable to itself represents a residual variance. Two latent variables are shown in circles for each factor of reading and mathematics. The reading latent variable has a path to each manifest variable (grade 1, 3 and 5). The variance of the reading construct was fixed at 1.0 to set the scale, which left three loadings and three residuals as freely estimated parameters of the model. Similarly, mathematics latent variable has a path to each manifest variable (grade 1, 3 and 5). The variance of mathematics construct was fixed at 1.0 to set the scale, which also left three loadings and three residuals as freely estimated parameters of the model.

Figure 1: Path diagram of latent variables of reading and mathematics with corresponding manifest variables and race, family, classroom, and school covariates



Limitations

Longitudinal studies cannot meet all of the rules for designing true experiments because age is a fixed personal attribute that cannot be experimentally assigned (Schaie, 2005). Consequently, longitudinal studies are subject to the types of problems associated with the kind of study called quasi experiments (Cook & Campbell, 1979). These problems may result to internal and external validity. Threats to validity alter accuracy in the interpretation of results in longitudinal studies. The threats to internal validity that beset longitudinal designs are maturation, regression to the mean, instrumentation effects (factorial invariance), selection effects, and confounding effects (age and time of measurement). The above threats to internal validity as they affect longitudinal studies and this dissertation are discussed as follow.

Maturation is the threat when an observed effect might be due to the respondent's getting older, wiser, stronger, and more experienced (Cook & Campbell, 1979). This study focuses on

Black-White test score differentials from kindergarten through eighth grade. Maturation is definitely not a threat to developmental studies but rather a primary variable of interest (Schaie, 2005). Maturation could also be a variable of interest in educational research because the hypothesis that the performance of a student on a test is function of age (maturation) is not implausible. As people mature so also does their knowledge in specific subject area and skills in test-taking. Therefore, an eighth grader should perform better than a kindergartener on a test that has common specifications.

Another important threat in educational research is regression to the mean. Regression to the mean involves the tendency of the variables containing measurement error to regress toward population mean from one occasion to the next (Schaie, 2005). This threat is a concern relative to a two occasion longitudinal designs; however, after the second occasion, regression effects are dissipated with no expected change (Nesselroade, Stigler, & Baltes, 1980). For this study, analyses will be conducted using six waves of data collection, and therefore regression to the mean as a threat will not be of significant concern.

The internal validity threat of instrumentation refers to differences in measurement techniques that covary with measurement occasions (Schaie, 2005). Here, three factors are important in addressing this threat: 1) lower completion rates from the previous data collection, 2) as children get older, they could refuse to cooperate at a much higher rate than younger children, and 3) changes in the field procedure relative to parent consent before the children could be approached. All these factors related to maturational trends may also in some way obscure reliability.

The threat of test-retest is when an effect might be due to the number of particular responses that are measured (Cook & Campbell, 1979). Specifically, familiarity with the test can enhance performance because responses and errors from previous testing occasions are likely to be remembered. In this study, with six data points, a high value of test-retest effects is not highly unlikely. Selection is a threat when an effect may be due to the process of obtaining a sample from the population such that the observed effect (test outcomes) arises from the specific sample characteristics rather than from the maturational effect to be estimated (Schaie, 2005). It is frequently impossible to rule out the possibility that differences across groups may be a function of differential recruitment (Schaie, 1959).

It is imperative to conclude that Black-White test score differential may be a function of multidimensional characteristics that shape both the social and cognitive developments of the child. In a longitudinal study, participants may drop out and be unable to fulfill the study requirements. This phenomenon is often referred to as attribution and may be as a result of death or cross-district movement.

4 Results

The results of this dissertation are presented in four sections. The first section gives the descriptive statistics of variables. The second gives the results that test hypotheses 1 and 2 using a 2 x 3 two-way repeated measures ANOVA to evaluate the magnitudes of test score differentials between Black and White students that increase from first through fifth grade. The scores used in the analysis were based on the assessments given to students in the spring of first, third, and fifth grades. The third section establishes factorial invariance of the latent constructs (i.e. reading and mathematics) across White and Black groups. Finally, section four shows the results of SEM using multiple indicators multiple causes (MIMIC) models (Jöreskog & Goldberger, 1975) to test hypotheses 3 and 4. This section shows the mediated effects of family, classroom (or teacher) and school covariates on Black-White test score differentials.

Descriptive Statistics

An inspection of the descriptive statistics from SPSS outputs are shown for both reading and mathematics as indicated in the following sections.

Reading

Table 9 shows descriptive statistics of reading output for three time points (grade 1, 3 and 5). The table shows that White group mean score was higher than that of the Black group at all three time points being considered. In spring first grade, the results indicate the White group mean was 12.26 points higher than their Black counterparts. In the spring of third grade, the results show White group scored 22.55 points higher than the Black group, and the White group scored 22.00 points higher than the Black in fifth grade. The score differentials between the White and Black group increased in reading as students advance from first through third grade. However, there was a slight decrease in the score differential as students advance from third grade through the fifth. This is in agreement with previous research studies that conclude that shrinkage occurs in the achievement gap as children advance from one grade level to another. Table 9 presents weighted reading scale score means for each round. These scores are estimates of the number of correct answers that would have been expected if at every round each child had been given all of the 212 test items (Najarian et al., 2009). This dissertation used three rounds of assessments, which were spring first, third and fifth grades because they were administered at equal intervals.

Table 9: Reading assessment IRT scale score means and standard deviations, rounds 4 through 6: School years 1999-2000, 2001-02, and 2003-04

		Race	
Variable	Full sample	White	Black
Spring first grade reading IRT score	80.40(22.99)	82.34(22.98)	70.14(20.16)
Spring third grade reading IRT score	131.64(26.69)	135.23(22.44)	112.68(25.12)
Spring fifth grade reading IRT score	154.19(24.93)	157.69(23.19)	135.69(25.67)

Notes: Table shows means and standard deviations (in parentheses) of reading scale scores from spring first grade through spring of fifth. Estimates are based on cross-sectional weights within each round. Full sample is based on the population of Black and White, N = 7430 (50.3% missing data). N for Black = 1182 and N for White = 6248. Estimates for kindergarten through eighth grade have been put on a common scale to support comparisons. The range of values is 0 – 212.

Mathematics

Table 10 shows a descriptive statistics of mathematics output for three time points (grades 1, 3 and 5). It indicates that White group mean score was higher than that of the Black group in all three time points being considered. In first grade, results indicate White group mean that was 13.83 points higher than their Black counterparts. In third grade, results show a White group mean that was 21.95 points higher than the Black group; the White group scored 23.33 points higher than the Black in fifth grade. As students advance from first grade through fifth, there was an increase in the magnitude of Black-White score differentials. Table 10 presents weighted mathematics scale score means for each round. These scores are estimates of the number of correct answers that would have been expected if at every round each child had been given all of the 174 test items (Najarian et al., 2009). This dissertation used three rounds of assessments – spring first, third and fifth grades because they were administered at equal intervals.

Table 10: Mathematics assessment scale score means, standard deviations, spring first grade through spring fifth grade: School years 1999-2000, 2001-02 and 2003-04

		Race	
Variable	Full sample	White	Black
First grade mathematics IRT score	64.41(17.99)	66.63(17.83)	52.80(13.92)
Third grade mathematics IRT score	102.15(24.00)	105.66(22.78)	83.71(21.69)
Fifth grade mathematics IRT score	125.94(24.07)	129.68(22.08)	106.35(24.57)

Notes: Table shows means and standard deviations (in parentheses) of mathematics scale scores from spring first grade through spring of fifth. Estimates are based on cross-sectional weights within each round. Full sample is based on the population of Black and White, N = 7466 (50.3% missing data). Estimates for kindergarten through eighth grade have been put on a common scale to support comparisons. The range of values is 0 – 174.

Two-way Repeated Measures ANOVA

In the following sections, a 2 x 3 two-way repeated measures ANOVA was conducted to investigate Black-White test score differentials in reading and mathematics to test hypotheses 1 and 2. The main grade effect and the grade x race interaction effect were tested using multivariate criterion of Wilk's lambda (Λ). Results of the reading and mathematics analyses will follow.

Reading

This study was conducted to evaluate test score differential between Black and White in reading. First, IRT reading scores for grades 1, 3 and 5 were analyzed by means of a 2 x 3 two-way repeated measures (or mixed design) ANOVA having two levels of race (White, Black) as a between-subjects factor and three levels of grade (grades 1, 3 and 5) as a within-subjects factor. A 2 x 3 repeated measures ANOVA was conducted between White and Black groups over three grade levels (grades 1, 3 and 5) examining differences in assessment scores. The grade main effect and the grade x race interaction effect were tested using multivariate criterion of Wilk's lambda (Λ). The multivariate tests indicate a significant grade main effect, $\Lambda = .13$, $F(2, 7427) = 25770.97$, $p < .001$, $\eta^2 = .87$. Statisticians indicate that an η^2 of .01 represents a small effect size, .06 a medium effect size, and .14 a large effect size (Green, Salkind, & Akey, 2000). Thus, the changes that occurred were significant and represented a relatively large effect size. Both Black and White students improved on their test scores from grade 1 to 3 and from grade 3 to 5; with most gains occurring from grade 1 to 3 (see Table 10 and Figure 2). In reading, as students advance from grade 1 to 3, Black-White gap widens, however, the gap narrows as they progress from grade 3 to 5. There was also a significant grade by race interaction, indicating that Black and White students have divergent scores over grades 1, 3 and 5

Because the main effect was significant, follow-up tests were conducted. Three paired-samples *t* tests were conducted to follow-up the significant main effect, controlling for familywise error rate across these tests by using a Holm's sequential Bonferroni approach. Differences in mean reading scores between White and Black students were significantly different between grades 1 and 5, $t(7571) = -321.32$, $p < .001$, between grades 1 and 3, $t(9646) = -255.50$, $p < .001$, and between grades 3 and 5, $t(7580) = -134.50$, $p < .001$. Although this design is nonorthogonal due to unequal sample sizes, the results of this analysis are meaningful and interpretable. The reasons for nonorthogonality may be classification factors, where unequal cell sizes reflect true differences in population sizes (Scott E Maxwell & Delaney, 2004). The full Black-White sample data used in this dissertation consisted of 7466 participants. Of these, 6270 (84.0%) were Whites and 1196 (16.0%) were Blacks.

The multivariate tests also indicate a significant grade-by-race interaction effect, $\Lambda = .96$, $F(2, 7427) = 159.94$, $p < .001$, $\eta^2 = .04$, reflecting a significant and small effect size. Figure 2 shows that White students ($M = 82.34$, $SD = 22.98$) scored higher than their Black counterparts ($M = 70.14$, $SD = 20.16$) in first grade. Also in third grade, White students ($M = 135.22$, $SD = 25.44$) scored higher than Black students ($M = 112.68$, $SD = 25.12$). Similarly, in fifth grade White students ($M = 157.69$, $SD = 23.19$) scored higher than Black students (135.69, $SD = 25.67$).

Because grade by race interaction was significant, follow-up tests were conducted. Three follow-up mixed ANOVAs tests were conducted for grades 1 and 3, grades 1 and 5, and grades 3 and 5. The first results of multivariate tests indicate a significant grade-by-race interaction effect between grade 1 and 3, $\Lambda = .96$, $F(1, 9645) = 400.28$, $p < .001$, $\eta^2 = .04$, reflecting a significant and small effect size. That is, White students ($M = 82.34$) scored higher in grade 1 reading assessment than their Black ($M = 70.14$) counterparts. White students ($M = 135.23$) also scored higher than Black students ($M = 112.68$) in grade 3. The second results of multivariate tests also indicate a significant grade-by-race interaction effect between grades 1 and 5, $\Lambda = .97$, $F(1, 7570) = 257.85$, $p < .001$, $\eta^2 = .03$, reflecting a significant and small effect size. The results further indicate that White students ($M = 82.34$) scored higher than Blacks ($M = 70.14$) in grades 1 and grade 5, White students ($M = 157.69$) also scored higher than Black students ($M = 135.69$). The third results of multivariate tests also indicate a nonsignificant grade-by-race interaction effect between grades 3 and 5, $\Lambda = 1.00$, $F(1, 7579) = 1.61$, $p = .21$, $\eta^2 = .00$, reflecting a nonsignificant and zero effect size. The results further indicate that White students ($M = 135.23$) scored higher than Blacks ($M = 112.68$) in grades 3 and in grade 5, White students ($M = 157.69$) also scored higher than Black students ($M = 135.69$). These results indicate that Black-White reading gap widens from grade 1 to 3, and narrows as students advance from grade 3 to 5.

Figure 2: Graphs showing grade by race interaction effect and changes over time for reading

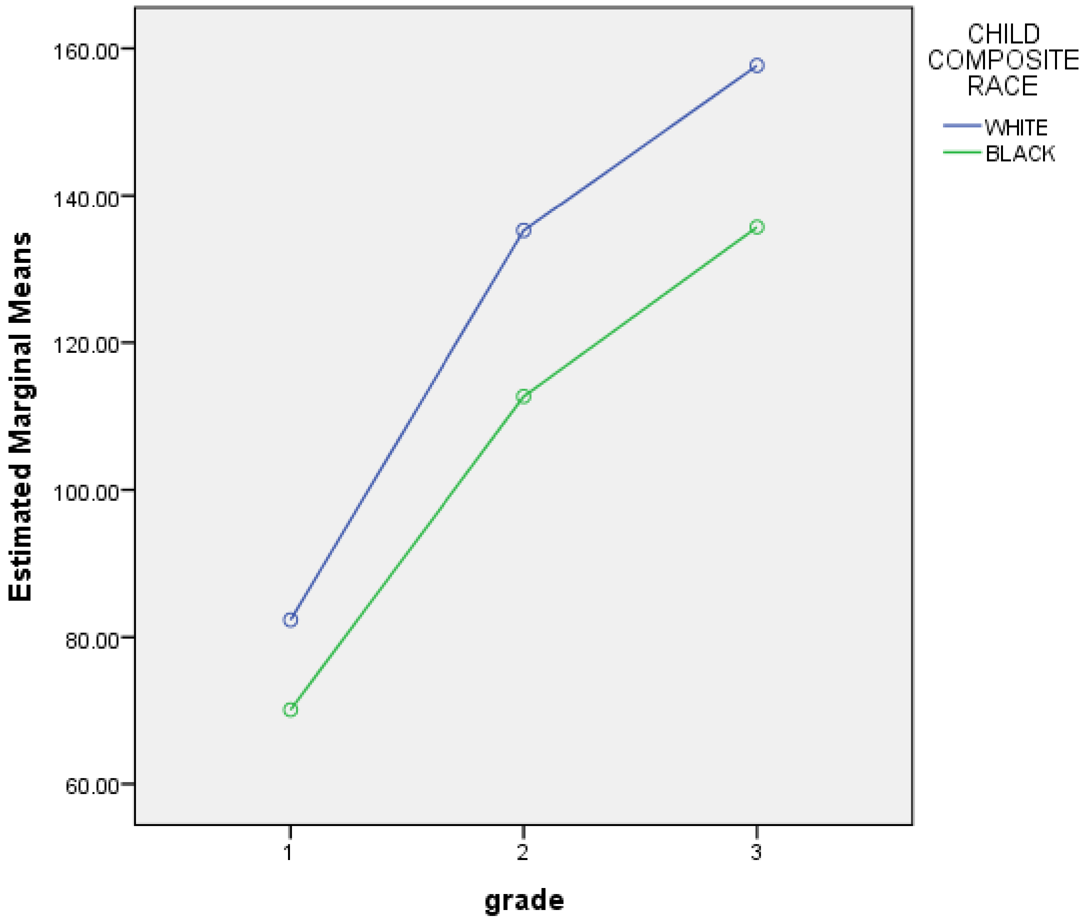


Table 11 shows the means, standard deviation and marginal means for race and reading scores for grades 1, 3 and 5. The table indicates that White students scored higher than their Black counterparts in readings from first grade through the fifth. Examining the interaction, White students started at mean score of 82.34 points in first grade, gained 52.89 points by the spring of third grade to obtain a mean score of 135.23 points, and increased by 22.46 points by spring of fifth grade, yielding a mean score of 157.69 points on that test. Black students started with a mean score of 70.14 points, scored 112.68 and 135.69 points in spring third and fifth grades, respectively. Further explication of the results indicate a Black-White score gap of 12.20, 22.55 and 22.00 points in first, third and fifth grades, respectively. These scores were based on IRT score estimates of 212 test items (R. G. Fryer & Levitt, 2004). Examination of the marginal means indicate that White students scored an average 125.09 points, while their Black counterparts scored 106.17 points, indicating a Black-White gap of an average of 18.92 points from first grade through fifth grade . These results indicate an increase in the magnitude of Black-White test score differential in reading widens as students advance from first to third grade and narrows from third to fifth grade.

Table 11: Means, standard deviations, marginal means, and mean score differentials for Black and White students in reading

Race	Grade						Marginal means
	First	Gap	Third	Gap	Fifth	Gap	
White	82.34 (22.98)		135.23 (25.44)		157.69 (23.19)		125.09
Black	70.14 (20.16)	12.20	112.68 (25.12)	22.55	135.69 (25.67)	22.00	106.17
Marginal means	76.24		123.96		146.69		115.63

Note: N for White group = 6270 and N for the Black group = 1196. Standard deviations are shown in parentheses. Marginal means are the means of the column scores and row scores.

Mathematics

This study was conducted to evaluate test score differential between Black and White in mathematics. To test this hypothesis, IRT mathematics scores for grades 1, 3 and 5 were analyzed by means of a 2 x 3 two-way repeated measures (or mixed design) ANOVA having two levels of race (White, Black) as a between-subjects factor and three levels of grade (grades 1, 3 and 5) as a within-subjects factor. A 2 x 3 repeated measures ANOVA was conducted between White and Black groups over three grade levels (grades 1, 3 and 5) examining differences in assessment scores. The grade main effect and the grade x race interaction effect were tested using multivariate criterion of Wilk's lambda (Λ). The multivariate tests indicate a significant grade main effect, $\Lambda = .12$, $F(2, 7463) = 26415.00$, $p < .001$, $\eta^2 = .88$. Statisticians indicate that an η^2 of .01 represents a small effect size, .06 a medium effect size, and .14 a large effect size (Green, Salkind, & Akey, 2000). Thus, the changes that occurred were significant and represent a relatively large effect size. Both Black and White students improved on their test scores from grade 1 to 3 and from grade 3 to 5; with most gains occurring from grade 1 to 3 (see Table 12 and Figure 2). In mathematics, as students advance from grade 1 to 3, Black-White gap widens, however, the gap narrows as they progress from grade 3 to 5. There was also a significant grade by race interaction, indicating that Black and White students have divergent scores over grades 1, 3 and 5

Because the main effect was significant, follow-up tests were conducted. Three paired-samples *t* tests were conducted to follow-up the significant main effect, controlling for familywise error rate across these tests by using a Holm's sequential Bonferroni approach. Differences in mean mathematics scores between White and Black students were significantly different between grades 1 and 5, $t(7576) = -324.72$, $p < .001$, between grades 1 and 3, $t(9700) = -241.81$, $p < .001$, and between grades 3 and 5, $t(7618) = -167.84$, $p < .001$. Although this design is nonorthogonal due to unequal sample sizes, the results of this analysis are meaningful and interpretable. The reasons for nonorthogonality may be classification factors, where unequal cell sizes reflect true differences in population sizes (Scott E Maxwell & Delaney, 2004). The full Black-White sample data used in this dissertation consisted of 7466 participants. Of these, 6270 (84.0%) were Whites and 1196 (16.0%) were Blacks.

The multivariate tests also indicate a significant grade-by-race interaction effect, $\Lambda = .95$, $F(2, 7463) = 192.58$, $p < .001$, $\eta^2 = .05$, reflecting a significant and small effect size. Figure 2 shows that White students ($M = 66.63$, $SD = 17.83$) scored higher than their Black counterparts

(M = 52.80, SD = 13.92) in first grade. Also in third grade, White students (M = 105.66, SD = 22.78) scored higher than Black students (M = 83.71, SD = 21.69). Similarly, in fifth grade White students (M = 129.68, SD = 22.08) scored higher than Black students (106.35, SD = 24.57).

Because grade by race interaction was significant, follow-up tests were conducted. Three follow-up mixed ANOVA tests were conducted for grades 1 and 3, grades 1 and 5, and grades 3 and 5. The first results of multivariate tests indicate a significant grade-by-race interaction effect between grade 1 and 3, $\Lambda = .96$, $F(1, 9699) = 421.50$, $p < .001$, $\eta^2 = .04$, reflecting a significant and small effect size. That is, White students (M = 66.63) scored higher in grade 1 mathematics assessment than their Black (M = 52.80) counterparts. In grade 3, White students (M = 105.66) scored higher than Black students (M = 83.71), while in grade 5 White students (M = 129.68) scored higher than Black students (M = 106.35), as well. The second results of multivariate tests also indicate a significant grade-by-race interaction effect between grades 1 and 5, $\Lambda = .96$, $F(1, 7575) = 350.90$, $p < .001$, $\eta^2 = .04$, reflecting a significant and small effect size. The results further indicate that White students (M = 66.63) scored higher than Black (M = 52.80) in grade 1 and in grade 5, White students (M = 129.68) also scored higher than Black students (M = 106.35). The third results of multivariate tests also indicate a significant grade-by-race interaction effect between grades 3 and 5, $\Lambda = 1.00$, $F(1, 7617) = 14.12$, $p < .001$, $\eta^2 = .00$, reflecting a significant but zero effect size. The results further indicate that White students (M = 105.66) scored higher than Blacks (M = 83.71) in grade 3 and in grade 5, White students (M = 129.68) also scored higher than Black students (M = 106.35). These results also indicate that Black-White mathematics gap widens from grade 1 to 3, and narrows as students advance from grade 3 to 5.

Figure 3: Graphs showing grade by race interaction effect and changes over time for mathematics

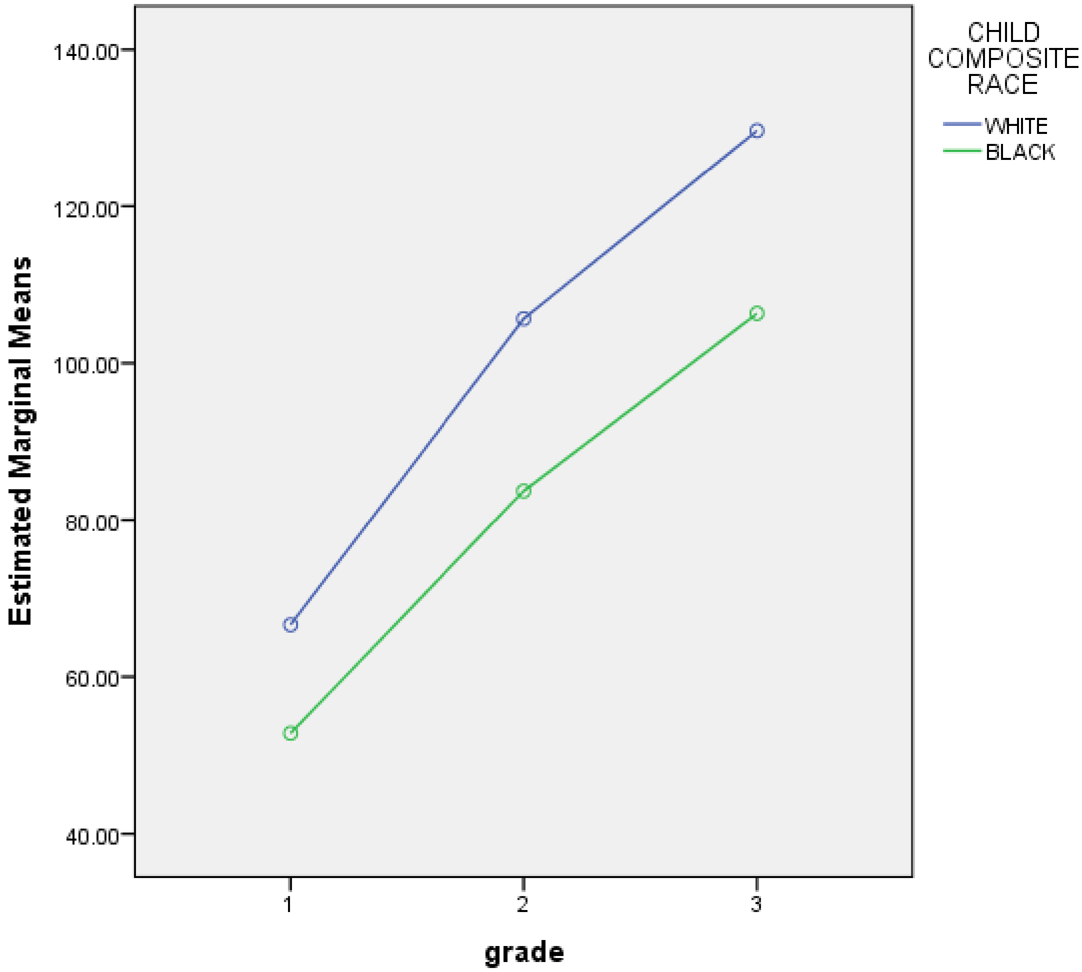


Table 12 shows the means, standard deviation and marginal means for race and reading scores for grades 1, 3 and 5. The table indicates that White students scored higher than their Black counterparts in readings from first grade through the fifth. Examining the interaction, White students started at a mean score of 66.63 points in first grade, gained 39.03 points by the spring of third grade to obtain a mean score of 105.66 points, and increased by 24.02 points by spring of fifth grade, yielding a mean score of 129.68 points on that test. While Black students started with a mean score of 52.80 points, scored 83.71 and 106.35 points in spring third and fifth grades, respectively. Further explication of the results indicate a Black-White score gap of 13.83, 21.95 and 23.23 points in first, third and fifth grades, respectively. These scores were based on IRT score estimates of 212 test items (R. G. Fryer & Levitt, 2004). Examination of the marginal means indicate that White students scored an average 100.66 points, while their Black counterparts scored 80.95 points. These results indicate an increase in the magnitude of Black-White test score differential in reading widens as students advanced from first to the third grade and narrows from third to fifth grade.

Table 12: Means, standard deviations, marginal means, and test score differentials for Black and White students in mathematics

	Grade						Marginal means
Race	First	Gap	Third	Gap	Fifth	Gap	
White	66.63		105.66		129.68		
	(17.83)		(22.78)		(22.08)		100.66
Black	52.80		83.71		106.35		
	(14.48)	13.83	(21.69)	21.95	(24.57)	23.33	80.95
Marginal means	59.72		94.69		118.02		90.81

Note: N for White group = 6270 and N for the Black group = 1196. Standard deviations are shown in parentheses. Marginal means are the means of the column scores and row scores.

Measurement invariance testing in MIMIC modeling

Before modeling the effects of each covariate on the latent constructs of reading and mathematics, levels of measurement invariance were conducted to test for equivalence of latent constructs across race or time points. To identify a noninvariant variable, a direct path from a grouping variable to an observed variable is tested in the model (Kim, Yoon, & Lee, 2012). The model with a direct path from a grouping variable to a measured variable can be written as

$$Y_{ij} = \lambda_j\eta_i + \beta_j X_i + \varepsilon_{ij}$$

$$\eta_i = \gamma X_i + \xi_i \quad (5)$$

where β_j is a path coefficient of the grouping variable in relation to the *j*th observed variable (Finch, 2005; Kaplan, 2009). To test for measurement invariance, this dissertation was conducted to establish the equivalence of indicator intercepts and factor means across groups of race with reference to MIMIC model. The steps involved in testing for measurement invariance in MIMIC modeling are two-fold. The first step is to regress the indicators (reading grades 1, 3, and 5; and mathematics grades 1, 3 and 5) onto the covariate race to test for the equivalence of indicator intercepts in both groups of race. The second step is to regress the latent variables (reading and mathematics) onto the covariate race to test for the equivalence of factor means across groups. Appendices 2 and 3 show M*plus* syntaxes to test for indicator intercepts and factor means, respectively.

Table 13 shows results of the tests of measurement invariance across groups. Model 1 tests for the equivalence of indicator intercepts across race by regressing indicators (grades 1, 3 and 5) onto race; while Model 2 tests for equivalence of factor means by regressing the latent variables onto race. The table's interpretation follows various absolute and incremental fit indices recommended by many researchers for assessing model fit. Hu and Bentler (1999) suggest the following guidelines between the target model and observed data. They suggest a reasonably good fit when 1) standardized root mean square residual (SRMR) values are .08 or below, 2) root mean square error of approximation (RMSEA) values are .06 or below, and 3) comparative fit index (CFI) and Tucker-Lewis index (TLI) values are .95 or greater. Browne and Cudeck (1993) propose that RMSEA values less than .08 suggest adequate model fit and RMSEA values less than .05 suggest a good model fit. Bentler (1990) also suggested that CFI and TLI in the range of 90-95 may be indicative of acceptable model fit. The overall fit of Model 1 (indicator intercepts) is acceptable, $\chi^2_{(8,\, n\, =\, 15012)} = 1299.8$, $p < .001$, SRMR = .026, RMSEA =

.104 (90% CI = .099 to .106), TLI = .93, CFI = .97. Similarly, the overall fit of Model 2 (factor means) is also acceptable, $\chi^2_{(21, n = 15012)}$ = 1547.7, p < .001 , SRMR = .026, RMSEA = .092 (90% CI = .088 to .096), TLI = .94, CFI = .97.

Table 13: Measurement invariance of a MIMIC model of reading and mathematics (N = 15012 imputed)

Model	χ^2	*Df*	CFI	TLI	SRMR	RMSEA (90% CI)
Mode 1	1299.8	8	.97	.93	.026	.104(.099-.109)
Model 2	1546.7	12	.97	.94	.026	.092(.088-.096)

Note: Model 1 tests for equivalence of indicator intercepts across groups by regressing race onto indicators; Model 2 tests for equivalence of factor means groups by regressing race onto the latent variables. CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; 90% CI = 90 percent confidence interval.

MIMIC models as an alternative to multiple-group analysis

MIMIC models was used in this dissertation as another way to estimate group differences between Black and White students on latent variables of reading and mathematics; where factors with effect indicators, grades 1, 3 and 5 were regressed onto the dichotomous cause indicator, race, that represent group membership. The total sample (N = 15012 imputed) was not partitioned into subsamples of Black and White groups. Therefore, there were no special identification requirements in M*plus* syntax beyond the single-sample analyses.

The MIMIC model in Figure 5 is a dichotomy that has a single-cause indicator that represent the race coded as 0 = White and 1 = Black. The reading factor has three effect indicators, grades 1, 3 and 5; and mathematics also with three effect indicators, grades 1, 3 and 5. In MIMC model, reading and mathematics factors are endogenous and thus has disturbances (error terms). Endogenousity indicates that the other variables in the model exert direct effects on both reading and mathematics factors. The MIMIC model did not have a mean structure; instead it had a covariance structure. This implies that all means were assumed to be zero; that is they were not analyzed. This study present the overall model fit and point-biserial correlations between the race variable and latent factors for each model. It also reports the unstandardized and standardized coefficients for direct effect of race covariate on reading and mathematics factors. Because race was coded as 0 = White and 1 = Black, positive regression weights indicate higher predicted overall standings on both reading and mathematics factors for Blacks than Whites. Conversely, negative regression weights indicate higher predicted overall standings on both reading and mathematics factors for Whites than Blacks. MIMIC models were used in this dissertation because they have the advantages of smaller size requirements, many groups comparisons, and more parsimonious because measurement model parameters are not estimated in each of the groups.

Structural equation modeling using MIMIC models

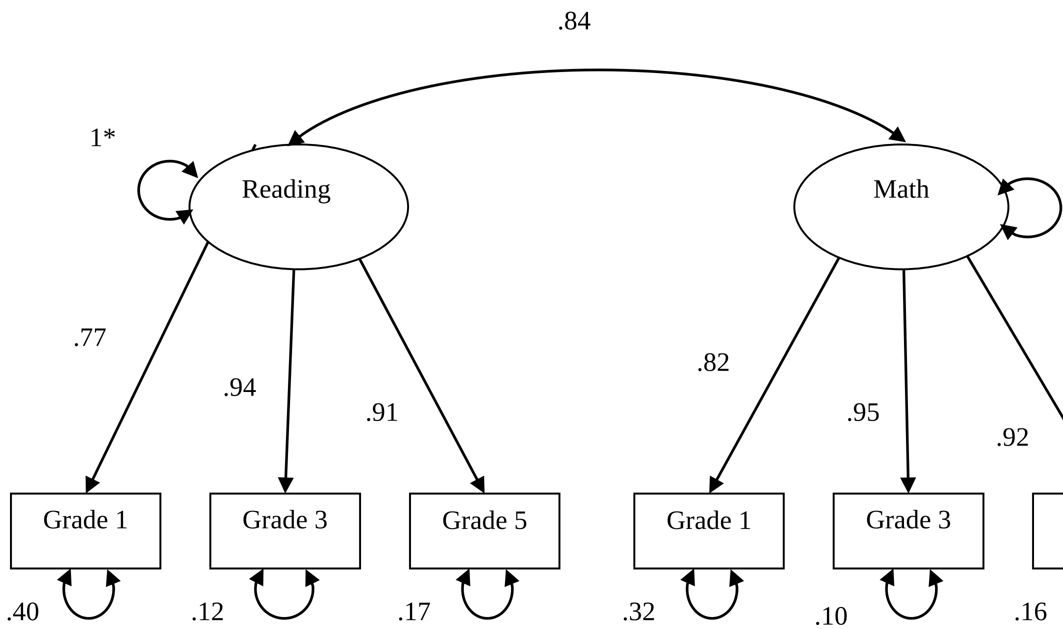
The approach MIMIC models, also known as CFA with covariates (Jöreskog & Goldberger, 1975) was used in this section. The MIMIC modeling approach used in this dissertation involved three basic steps. The first step was to ensure that the two-factor CFA model of reading and mathematics was reasonable and good fitting in the full sample (N = 15012 imputed). In the second step, covariates were added to the baseline model to evaluate the effects of covariates on the latent constructs of reading and mathematics. In this step, analyses were conducted in 2 stages, Model 2 and 3. In Model 2, race covariate was added to the two-factor CFA model. In Model 3, family, classroom and school covariates were added to Model 2. The last step was culminated by the interpretation of the results of each model. This step includes the interpretation of reading and mathematics factors and the effect of covariates on these factors.

MIMIC modeling approach was used in this dissertation to study the effects of family, classroom, and school on White-Black test score differentials in hypotheses 3 and 4. Family factors included mother education, father education, parental education, father occupational prestige, mother occupational prestige, level of poverty, the number of books child has in the house, child visits to the library, and child home computer for use. Classroom factors encompassed class enrolment, teacher experience, and highest level of education obtained by the teacher; while school covariates included school enrolment and, free and reduced lunch.

Model 1 was to ensure that the two-factor CFA model of reading and mathematics was a reasonable and good fit in the full sample (N = 15012 imputed). In this step, race, family, classroom and school covariates were not included in the CFA model. The variances of the factors were fixed at 1.0 to set the scale, leaving the covariance of the two factors and the eight loadings and eight residuals as freely estimated. Figure 5 shows a two-factor CFA model while Appendix 3 shows the M*plus* syntax for the model. Model 1 provided a good fit to the data, $\chi^2(8) = 1345.8$, $p < .001$, CFI = .97, TLI = .94, RMSEA = .106 (90% CI = .101 to .110), SRMR = .026. All parameter estimates were reasonable and statistically significant – as the completely standardized factor loadings range from .77 to .95 and correlation between reading and mathematics is .84.

In this dissertation, three reading indicators (grades 1, 3 and 5) and three mathematics indicators (grade 1, 3 and 5) were used with an imputed sample size of N = 15012. As shown in Table 14, the three indicators of reading and three indicators of mathematics were strongly intercorrelated. It was conjectured that each of these IRT scale scores, grades 1, 3 and 5 were manifest indicators of the latent constructs of reading and mathematics. That is, each of the observed indicators had shared reading and mathematics abilities that accounted for the intercorrelations among these observed measures. The indicators were correlated because they shared common academic skill development in reading and mathematics. However, if these latent constructs were partialed out, no relationship would be seen.

Figure 4: Two-factor confirmatory analysis of reading and mathematics



Model Fit: $\chi^2_{(8,\ n=15012)}$ = 1345.8; RMSEA = .106$_{(.101;\ .110)}$; CFI = .97; TLI = .94

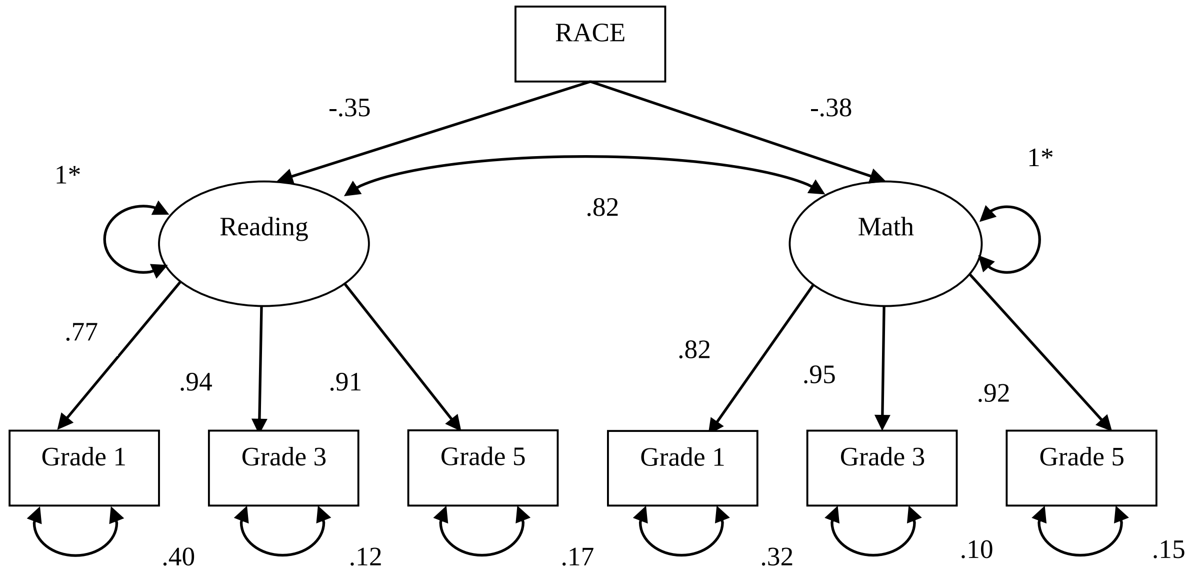
Table 14: Intercorrelations among reading and mathematics IRT scale scores for first grade through fifth grade

	ReadT4	ReadT5	ReadT6	MathT4	MathT5	MathT6
ReadT4	1.000					
ReadT5	0.729	1.000				
ReadT6	0.686	0.856	1.000			
MathT4	0.663	0.649	0.636	1.000		
MathT5	0.631	0.741	0.714	0.785	1.000	
MathT6	0.594	0.712	0.742	0.742	0.874	1.000

In Model 2, race was added to the two-factor CFA model as a covariate. Appendix 5 provides M*plus* syntax for this model specification. As shown in Figure 6, reading and mathematics latent factors were regressed onto race covariate using the “ON” keyword. In this model, reading indicators at grades 1, 3, and 5 were regressed onto the reading factor and mathematics indicators at grades 1, 3 and 5 onto the mathematics factor. Table 15 shows that both the reading and mathematics indicators were strongly correlated, however, there were weak correlations between race and the manifest variables. Additional results indicated 12.2% of the variability can be accounted for in the reading construct while 14.4% of the variability can be accounted for by the mathematics construct. Model 2 with race as a covariate provided a good fit to the data, $\chi^2(12) = 1546.7$, $p < .001$, CFI = .97, TLI = .94, RMSEA = .092 (90% CI = .088 to .096). Inclusion of the race covariate did not alter the factor structure.

Table 15 shows results of unstandardized and standardized solutions. Of particular interest are the regressive paths linking race to the latent factors. The paths of race to reading (z = -44.13, $p < .001$) and mathematics (z = -50.83, $p < .001$) were statistically significant. Given how the race covariate was coded (0 = White, 1 = Black) and the negative sign of the parameter estimate, reading unstandardized estimate = -0.91 indicates that Black students has a lower mean score than their White counterparts on the reading factor. Specifically, the mean of White students in reading is 0.91 units higher than the mean of Black students. With respect to mathematics, the unstandardized estimate of -1.00 indicates that Black students have a lower mean score than their White counterparts on the mathematics factor. Specifically, the mean of White students in mathematics is 1.00 units higher than the mean of Black students. Additional results indicates that the correlation between race and the reading factor is -.35; while correlation between race and the mathematics factor is -.38, indicating that race has little effect on the Black-White achievement gap.

Figure 5: Path diagram of MIMIC model of reading and mathematics as latent constructs and race as a covariate



Model Fit: $\chi^2_{(12,\ N = 15012)}$ = 1546.7; RMSEA = .092$_{(.088;\ .096)}$; CFI = .97; TLI = .94

Table 15: Results of baseline MIMIC model of reading and mathematics with race as a covariate

	Reading		Mathematics	
Model parameter	Unstandardized estimate/z-score	Standardized estimate/z-score	Unstandardized estimate/z-score	Standardized estimate/z-score
Race	-.909(-39.790)***	-.348(-44.129)***	-.999(-44.939)***	-.380(-50.832)***

Note: Table shows results based on 10 sets of imputed data of full sample N = 15,012. ***p < 001

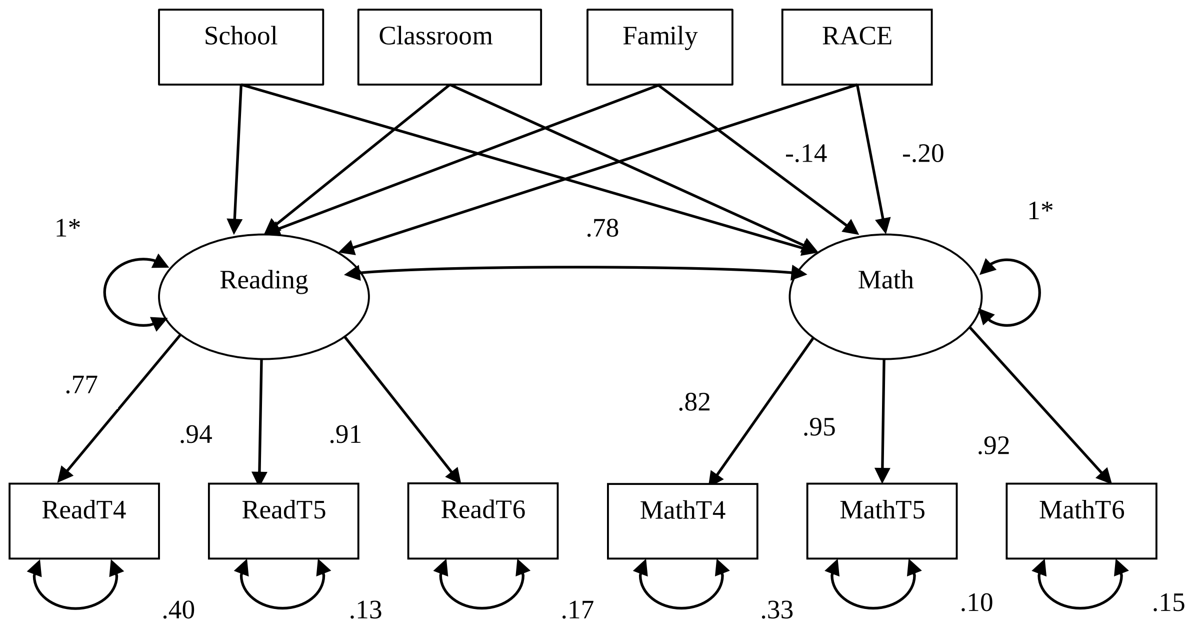
Table 16: Intercorrelations among reading and mathematics IRT scale scores at first grade through fifth grade and race as a covariate

	ReadT4	ReadT5	ReadT6	MathT4	MathT5	MathT6	Race
ReadT4	1.000						
ReadT5	0.729	1.000					
ReadT6	0.686	0.856	1.000				
MathT4	0.663	0.649	0.636	1.000			
MathT5	0.631	0.741	0.714	0.785	1.000		
MathT6	0.594	0.712	0.742	0.742	0.874	1.000	
Race	-0.211	-0.325	-0.340	-0.291	-0350	-0.374	1.000

Finally, in Model 3 family, classroom and school characteristics were added to Model 2 as shown in Figure 6 to evaluate the effects of covariates on the Black-White test score divergence. This model provided a good fit to the data, $\chi^2(100) = 387.8$, $p = 1.00$, CFI = .98, TLI = .97, RMSEA = .015 (90% CI = .014 to .017). Results indicate that the paths of race to reading and mathematics were statistically significant ($z = -13.48$, $p < .001$ and $z = 16.87$, $p < .001$, respectively). Additional results indicate that given how the race covariate was coded (0 = White, 1 = Black) and the negative sign of the parameter estimate, reading unstandardized estimate = -.40 indicates that Black students has a lower mean score than their White counterparts on the reading factor. Specifically, the mean of White students in reading is .40 units higher than the mean of Black students. Similarly, in mathematics the unstandardized estimate of -.58 indicates that Black students has a lower mean score than their White counterparts on the mathematics factor. Specifically, the mean of White students in mathematics is .58 units higher than the mean of Black students.

Additional results indicate that 31.5% of the variability can be accounted for in reading construct while 31.0% of the variability can be accounted for by the mathematics construct. With the addition of family, classroom and school characteristics to race, there is a decrease in score differentials between White and Black students. There is a decrease from .91 to .40 units and 1.00 to .58 units in reading and mathematics, respectively. Additional results indicate that the correlation between race and the reading factor is -.14; while correlation between race and the mathematics factor is -.20, indicating that family, classroom and school covariates has greater effect than race on the Black-White achievement gap. That is, covariates has a significant effect on closing test score differentials between White and Black students.

Figure 6: Path diagram of MIMIC model showing reading and mathematics as latent constructs and race, family, classroom, and school as covariates



Model Fit: $\chi^2_{(100,\ N = 15012)}$ = 337.3; RMSEA = .013$_{(.011;\ .014)}$; CFI = .98; NNFI = .98

Summary of fit indices for CFA and MIMIC models

Table 17 shows summary fit indices for the baseline CFA model, CFA with race covariate, and CFA with race, family, classroom and school covariates models. Model fit values indicate that as family, classroom and school factors were added to the model, improvements were made to the fit indices between the target model and the observed data. That is, with the addition of these covariates to the baseline CFA model, there was a shrinkage in the Black-White test score differentials.

Results indicate that the CFA and the two MIMIC models provided good fit to the data. Models assessments were based on various absolute and incremental fit indices (Bentler & Bonett, 1980; Cheung & Rensvold, 2001; Ding, Velicer, & Harlow, 1995; Hu & Bentler, 1999). Model 1 provided acceptable fit to the data 1, $\chi^2_{(8,\ n=15012)}$ = 1345.8; RMSEA = .106$_{(.101;\ .110)}$; CFI = .97; TLI = .94. Model 2 comprising of race as a covariate also provided good fit to the data,

$\chi^2_{(12,\ n=15012)}$ = 1546.7; RMSEA = $.092_{(.088;\ .096)}$; CFI = .98; TLI = .97. Finally, Model 3 where all the covariates were added also provided good fit to the data, $\chi^2_{(100,\ n=15012)}$ = 337.3; RMSEA = $.013_{(.011;\ .014)}$; CFI = .98; TLI = .98. All the models were statistically significant at the $p < .001$. As race, family, classroom and school covariates were added to the CFA model, there was an improvement in model fit compared to the model that preceded it. These results reaffirm the interpretation of Black-White test score differentials using the unstandardized solutions.

Table 17: Fit statistics of MIMIC models with race, family, classroom, and school covariates

Model	χ^2	*df*	CFI	TLI	SRMR	RMSEA(90%CI)
Model 1	1345.8***	8	.97	.94	.026	.106(.101-.110)
Model 2	1546.7***	12	.97	.94	.026	.092(.088-.096)
Model 3	337.3***	100	.98	.98	.009	.013(.011-.014)

Note: Model 1: CFA without covariates; Model 2: Baseline MIMIC model with race as a covariate; Model 3: Model 2 plus family, classroom, and school covariates. CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; 90% CI = 90 percent confidence interval. $^{***}p < .001$

5 Discussion

The purposes of this dissertation are two-fold. The primary purpose was to investigate the magnitude of Black-White test score differentials in reading and mathematics as students advance from kindergarten through eighth grade using the analysis of variance (ANOVA) framework. The secondary purpose was to evaluate the effects of family, classroom (or teacher) and school on the test score divergence between the Black and White groups; and provide a better understanding of the size of the gap using the multiple indicator, multiple causes (MIMIC) modeling framework. The dissertation concludes with a discussion of the limitations of this research and directions for future research.

Summary of findings

The results of the study reaffirmed previous research that report that as students advance through grade levels, the magnitudes of the test score differentials increase (Fryer Jr & Levitt, 2004; Phillips et al., 1998) in both reading and mathematics. Analyses were conducted using Early Childhood Longitudinal Study Kindergarten (ECLS-K) 1998-99 consisted of seven waves of data collection from 21,260 students in 944 kindergarten programs. Of these, only three waves of data with equal time intervals were used. These were data points for spring first, third and fifth grades. To model the trajectory of the test score differentials, two different statistical designs were used. These were two-way repeated measures ANOVA and MIMIC (Jöreskog & Goldberger, 1975) models.

Results of the 2 x 3 two-way repeated measures ANOVA used in evaluating the mean differences between grade levels confirmed increases in test score differentials between Black and White students in both reading and mathematics as they advance from first grade through fifth grade. The ANOVA model results also concluded significant main grade effect and race x grade interaction effects in both reading and mathematics. Because of the significant main effect, paired-samples *t* tests were conducted to follow-up the significant main effect. All the three tests among the means for first, third and fifth grades were also significant, after controlling for Type 1 error across the three tests at the .05 level. Results of the ANOVA design also revealed that the test score differential in reading increased from 12.20 points in grade 1 to 22.55 units in grade 3 grade. From third grade to the fifth, the gap decreased from 22.55 to 22.00 points, indicating that as students progressed from third grade to the fifth, there was a shrinkage in the Black-White test score gap. This is in support of previous research studies that report that the Black-White gap shrinks as students advance through k-8 schooling. The ANOVA results for mathematics were similar to the reading. From first grade through the third, the mathematics gap increased from 13.83 points in first grade to 21.95 points in third grade. From third to fifth grade, results also showed a slight increase in the test score differential, from 21.95 points in third grade to 23.33 points in fifth grade.

The second analytical strategy, MIMIC models used in this dissertation to evaluate the effects of family, classroom (or teacher) and school on the Black-White gap revealed interesting results. Prior to modeling the effects of covariates, results of MIMIC factorial invariance showed a good fit to the data. That is, reading and mathematics constructs were factorially invariant across groups. MIMIC models were conducted with the addition of race, family, classroom and school covariates to the model. With the addition of covariates, results indicated statistically significant effect at the .05 level on the gap. Overall, the cumulative results of MIMIC models revealed that as different covariates (family, classroom and school) were added to the model, there were decreases in the Black-White test score differentials for both reading and mathematics.

Interpretation of ANOVA designs and MIMIC models

Reading and mathematics item response (IRT) scale scores were analyzed by means of two-way repeated measures ANOVA having two levels of race (Black, White) as a between-subjects factor and three levels of IRT scale scores (first, third and fifth grades) as within-subjects factor. The analysis was conducted to evaluate Black-White test score differentials. Interpretation of repeated measures ANOVA design is dependent on the multivariate criterion of Wilk's lambda (Λ), assuming multivariate normality and independence of observations hold (Scott E Maxwell & Delaney, 2004). The multivariate tests indicated that both main effect and race-by-grade interaction effect were statistically significant at the .05 level. That is, a significant test score differential existed between Black students and their White counterparts.

Interpretation of the MIMIC models is dependent upon whether or not reading and mathematics constructs were factortially invariant across Black and White groups. It is important to note that MIMIC models test only the invariance of indicator intercepts and factor means. Thus, MIMIC assumed that all other measurement and structural parameters are the same (Brown, 2012) across all levels of race. Two potential sources of MIMC factorial invariance, indicator intercepts and factor means were tested. The dissertation used four different fit indices to evaluate the reasonability of a good fit to the data. Many researchers have recommended model fit based on absolute and incremental fit indices (Bentler & Bonett, 1980; Cheung & Rensvold, 2001; Ding et al., 1995; Hu & Bentler, 1999). These fit indices were RMSEA, CFI, TLI and SRMR. Fit indices are guidelines in SEM parlance to determine if a particular model fits a data structure. Upon satisfactory determination of factorial invariance, MIMIC model analysis was conducted in two stages. The first stage established a confirmatory factor analysis (CFA) structure; while the second stage established the effects on family, classroom and school on Black-White gap. In this section, interpretation of results were based on both the unstandardized solutions and intercorrelations between variables. Race is a nominal variable that represents levels of known groups (0 = White, 1 = Black). A negative value of unstandardized estimate concluded that Black students have a lower mean than Whites on the reading and mathematics factors. Conversely, a positive value of unstandardized estimate concluded that Black students have a higher mean values than the White students.

Limitations and directions for future research

As with any study, this dissertation was limited by a number of factors. First, the study was limited by the data extracted from NCES. The license granted by NCES to use the data for this study only contained IRT scale scores for seven waves of data collection. Access was not given to the item-level scores to conduct factorial invariance using multiple groups. MIMIC could only examine two potential sources of invariance to test for factor means and indicator intercepts. Alternatively, access to item-level scores would have permitted the analysis of multiple-group CFA that should have tested all potential aspects of invariance. As a result of limited access given to the data used in this study, the researcher was faced with a large amount of missingness. As with any longitudinal studies in behavior research, an enormous amount of missingness due to attrition is not uncommon. In any longitudinal study, a participant may not partake in all the rounds of data collection due to reasons, such as relocation to another country or may voluntarily withdraws from the study.

To prevent loss of power due to missingness, multiple imputation mechanism was employed. As a result, M*plus* was unable to calculate modification indices that should have allowed researcher to evaluate direct and indirect effects of the manifest variables on the latent constructs of reading and mathematics. MIMIC models extend multigroup CFA (MGCFA) by analyzing the indirect effects of family, classroom, or school factors on latent reading and mathematics variability in order to further analyze measurement invariance and population heterogeneity (Muthén & Muthén, 2010). In a MIMC framework, an indirect effect for a given covariate would indicate that mean differences for latent reading and mathematics variability were found as a function of differing levels of that covariate.

With full access to ECLS-K 1998-99 data, including item-level data, future research should examine all aspects of invariance that should evaluate that the same latent constructs were being measured across groups of race. With this test, researchers should be able to determine the heterogeneity or differential item functioning (DIF) of the test items. In addition, M*plus* could easily calculate modification indices that could allow evaluation of direct and indirect effects of the manifest variables on the covariates. Future research should also examine White-Hispanics test score differentials in reading and mathematics. Also, future research should examine if other student behavior factors have any effect on the Black-White test score differentials.

Summary

This dissertation investigated the Black-White test score differentials in reading and mathematics from first grade through fifth grade. In addition, it investigated the effects of family, classroom (or teacher) and school on the test score gap. The dissertation contributes to research in the area of achievement gap between advantaged and disadvantaged groups. The major finding from this dissertation revealed test score differentials between Black and White students, and that the size of the gap increased as students advance through grade levels. This study seeks to further contribute to the literature by employing a MIMIC design to investigate the impact of family, classroom, and school covariate on test score differentials in reading and mathematics.

To date, researchers such as Brooks-Gunn et al (2003) have investigated the effect of family characteristics on Black-White test score gap using ordinary least square multiple linear regressions to estimate the mean difference in IQ and PPVT-R scores of Black and White children. Similarly, Fryer and Levitt (2004) have used ECKS-K data to investigate Black-White test score gap of children in the first two years of schooling. They presented a series of estimates of racial test score gap for the test taken in the fall of kindergarten. However, none of them have used MIMIC model to investigate test score gap between Black and White students using multiple time points.

Specifically, as family, classroom (or teacher) and school covariates were added to the models, the gap initially increased as students advanced from first grade to third grade; however, the gap decreased drastically or closed as students advanced from third to fifth grade. The magnitudes of the gap were based on IRT scores based on mean estimates of 212 and 174 test items for reading and mathematics, respectively. These findings were based on the limited data access to the ECLS-K 1998-99 longitudinal study.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*(4), 1012-1028.

Algina, J., & Crocker, L. (2006). *Introduction to classical and modern test theory*. Mason, OH: Thomson Wadsworth.

Allison, P. D. (2001). *Missing data* (Vol. 136). Thousand Oaks, CA: Sage Publications.

Ames, L. B. (1967). *Is your child in the wrong grade?* : Modern Learning Pr/Programs. New York, NY: Haper and Row.

Armor, D. (2003). *Maximizing intelligence*. New Brunswick, NJ: Transaction Publishers.

Bali, V. A., & Alvarez, R. M. (2004). The race gap in student achievement scores: Longitudinal evidence from a racially diverse school district. *The Policy Studies Journal*, *32*, 393-415.

Baron, R., Tom, D. Y. H., & Cooper, H. M. (1985). *Social class, race and teacher expectations. In Jerome B. Dusek, (Ed.), Teacher expectancies*, 251 - 269.

Behind, N. C. L. (2002). Act of 2001, Pub. L. No. 107-110, § 115. *Stat, 1425*, 107-110.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin, 88*(3), 588.

Borman, G., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's Equality of Educational Opportunity data. *The Teachers College Record, 112*(5), 1-2.

Bornstein, M. H., & Sigman, M. D. (1986). Continuity in mental development from infancy. *Child Development*, 251-274.

Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Brooks-Gunn, J., Duncan, G. K., & Klebanov, P. K. (1994). Economic deprivation and early-childhood development. *Child Development, 65*(2), 296-318.

Brooks-Gunn, J., Duncan, G. K., & Klebanov, P. K. (1995). Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. *Child Development, 67*(2), 396-408.

Brooks-Gunn, J., & Duncan, P. K. (1997). *The consequences of growing up poor*. New York, NY: Russell Sage Foundation.

Brooks-Gunn, J., & Duncan, P. K. (2000). Family poverty, welfare reform, and development *Child Development, 71*(1), 188-196.

Brooks-Gunn, J., Klebanov, P. K., Smith, J., Duncan, G. J., & Lee, K. (2003). The Black-White test score gap in young children: Contributions of test and family characteristics. *Applied Developmental Science, 7*(4), 239-252.

Brown, T. A. (2012). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W., Cudeck, R., & Bollen, K. A. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154*, 136-136.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology, 96*(1), 31.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (2000). NAEP Trends. *National Center for Education Statistics*, *2*(4), 31.

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*(3), 236-264.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Cole, D. A., & Maxwell, S. E. (2003). Testing meditational models with longitudinal data: Questions and tips in the use of structural equation modeling *Journal of Abnormal Psychology*, *112*, 558-577.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, F., Weinfield, F., & York, R. (1966). *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational Measurement (2nd Ed.) (pp. 443-507). Washington, D. C.: American Council on Education.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In: Collins, L.M.; Sayer, A.G.; editors. New methods for the analysis of change. Washington, DC. *American Psychological Association*, 107-135.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence Retrieved April 4, 2007 http://epaa.asu.edu/epaa/v8n1/.

Desimone, L., & Long, D. A. (2010). Teacher effects and the achievement gap: Do teacher and teaching quality influence the achievement gap between Black and White and high-and low-SES students in the early grades. *Teachers College Record, 112*(12), 3024-3073.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal, 2*(2), 119-143.

Downey, D. B., & Pribesh, S. (2004). When race matters: student/teacher racial matching and teachers' evaluations of students' behavior. . *Sociological Education, 77*, 267-282.

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year *American Sociological Review, 69*, 613-635.

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *The future of children*, *15*(1), 35-54.

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resourses account for racial and ethnic test score gaps? *Future of Children, 15*(1), 35-54.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Publications.

Fagan, J. F., & Singer, L. T. (1983). Infant recognition memory as a measure of intelligence. *Advances in infancy research*, 2, 31-78.

Ferguson, R. F. (1998). *Teachers' perceptions and expectations and the black–white test score gap. In C. Jencks & M. Phillips (Eds.), The black–white test score gap.* (pp. 273−317).

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-whitre test score gap. *Urban Education, 38*(4), 460-507.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis (2nd ed.).* Hoboken, NJ: Wiley and Sons.

Fryer Jr, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447-464.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics*, *86*(2), 447-464.

Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review, 8*(2), 249-281.

Gandara, P., Rumberger, R. W., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in Carlifornia schools: Unequal resources, unequal outcomes. Retrieved September 6, 2006, from http://epaa.asu.edu/epaa/v11n36/

Goldhaber, D. D. (2002). The mystery of good teaching: Surveying the evidence on student achievement and teachers' characteristics. *Education Next*, *2*(1), 50-55.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of psychology*. Hoboken, NJ: John Wiley and Sons.

Green, S. B., Salkind, N. J., & Akey, T. M. (2000). *Using SPSS for wimndows: Analyzing and understanding daaata.* Eaglewood Cliffs, NJ: Prentice-Hall.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for specialized populations: Does special education raise achievement for students with disabilities? *Review of Economics and Statistics, 84*(4), 584-599.

Haycock, K. (1998). Good teaching matters... a lot. *Magazine of History,* 61-63.

Herrnstein, R. J., & Murray, C. (1995). *The Bell Curve: Intelligence and class structure in American life.* New York, NY: Simon and Schuster.

Howell, P. L., & Miller, B. B. (1997). Sources of funding for schools. *The future of children*, 39-50.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Huck, S. W., Cormier, W. H., & Bounds, W. G. (1974). *Reading statistics and research*: Harper & Row New York.

IBM. (2010). IBM SPSS Statistics for Windows. Armonk, NY: IBM Corporation.

Ilg, F. L., Ames, L. B., Haines, J., & Gillespie, C. (1978). *School readiness: Behavior tests used at the Gesell Institute.(Rev ed)*: Harper & Row.

Jencks, C., & Phillips, M. (1998a). *The Black-White test score gap.* Washington, D.C.: Brookings.

Jencks, C., & Phillips, M. (1998d). *The black-white test score gap.* Washington, D.C.: Brookings

Johnson, C. M., & Bradley-Johnson, S. (2002). Construct stability of the cognitive abilities scale-for infants and toddlers. *Journal of Psychoeducational Assessment, 20*(2), 144-151.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351a), 631-639.

Kahlenberg, R. D. (2001). All together now: creating middle-class schools through public school choice. Washington, D. C.: Brookings.

Kantor, H., & Lowe, R. (2006). From New Deal to no deal: No Child Left Behind and the devolution of responsibility for equal opportunity. *Harvard Educational Review, 76*(4), 474-502.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (Vol. 10). Thousand Oaks, CA: Sage Publications.

Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of educational psychology, 101*(4), 765.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing Measurement Invariance Using MIMIC Likelihood Ratio Test With a Critical Value Adjustment. *Educational and Psychological Measurement*, *72*(3), 469-492.

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19*(6), 513-526.

Law, P. (2002). No child left behind act of 2001. *Public Law, 107*, 110.

Lee, V. E., & Bryk, A. (1989). A multilevel model of the social social distribution of high school acievement. . *Sociology of Education, 62*, 172-192.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*, 357.

Magnuson, K. A., & Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The future of children*, *15*(1), 169-196.

Matthews, L. L. (1996). *Academic and Adjustment Outcomes of Developmental Placement: A Longitudinal Study.* Albany, NY: State University of New York, Albany.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analysis of longitudinal mediation. *Psychological methods*, *12*, 23-44.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (Vol. 1). New York, NY: Psychology Press.

McArdle, J. J. (2001). *A latent difference score approach to longitudinal dynamic structural analyses* S. d. T. R. Cudeck, & D. Sorbom, Eds. (Ed.) *Structural Equation Modeling: Present and Future* (pp. 342-380).

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*(1), 110-133.

McArdle, J. J., & Hamagami, A. (2001). Latent difference scores structural models for linear dynamic analyses with incomplete longitudinal data. In: Collins, L. M; Sayer, A. G., Eds. New methods for the analysis of change. 139-175.

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development, 64*(1), 57-79.

Meredith, W., & Tisak, J. (1984). Statistical considerations in Tuckerizing curves with emphasis on growth curves and cohort sequential analysis. In annual meeting of the Psychometric Society.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrica, 55*, 105-122.

Muthén, L. K., & Muthén, B. O. (2010). Mplus (Version 6). Los Angeles, CA: Muthén & Muthén.

Najarian, M., Pollack, J. M., Sorongon, A. G., & Hausken, E. G. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K): Psychometric Report for the Eighth Grade: Washington, DC: National Center for Education Statistics.

NCES. (2000). *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance*. (Report 2000469). Washington, D.C.: National Center for Education Statistics

NCES. (2001). *The Condition of Education 2001*. (Report 2001072). Washington, D.C.: National Center for Education Statistics.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression towards the mean and the study of change. *Psychological bulletin, 88*, 622-637.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. *Real data analysis*, 31-78.

Phillips, M., Crouse, J., & Ralph, J. (1998). "Does Black-White Test Score Gap Widen After Children Enter School?" In the Black-White Test Score Gap, C. Jencks and M. Phillips, eds. (pp. 229-272). Washington, DC: The Brookings Institute.

Piaget, J. (1950). The Psychology of Intelligence.(translated by M. Piercy and EE Berlin). New York, NY: Routledge Publishing.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent Growth Curve Modeling.* Thousand Oaks, CA: Sage Publications.

Raver, C. C., Aber, L., & Gershoff, E.T. (2007). Testing Equivalence of Mediating Models of Income, Parenting, and School Readiness for White, Black and Hispanic Children in a National Sample. *Child development, 78*(1), 96-115.

Reardon, S. F. (2003). *Sources of educational inequality: the growth of racial/ethnic and socioeconomic score gaps in kindergarten and first grade.* Population Research Institute Working Paper #2003-05R. Pennyslavania State University, University Park, PA.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (1998). *Teachers, schools, and academic achievement.* Cambridge, MA: National Bureau of Economic Research

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin, 90*, 726-748.

Roscigno, V. J. (1998). Race and the reproduction of educational disadvantage. *Social Forces, 76*, 1033-1061.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel psychology, 47*(3), 537-560.

Rothstein, R. (2004). *Class and schools: Using social economic, and educational reform to close the black black-white achievement gap.* New York, NY: Teacher's College, Columbia University.

Rumberger, R. W., & Palardy, G. J. (2004). Multilevel study of school effectiveness research. In D. Kaplan (Ed.), Handbook of Quantitative Methodology for Social Sciences (pp. 235-258). Thousand Oaks, CA: Sage Publications.

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student compodsition on academic achievement in high school. . *Teachers College Record, 107*(9), 1999-2045.

Sanders, W. L. (1998). Value-added assessment. *The School Administrator, 55*(11), 24-32.

Sanders, W. L., & Rivers, J. C. (1996). Research progress report: Cumulative and residual effects of teachers on futrure student academic achievement. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Scarr, S. (1981). *Race, social class, and individual differences in I.Q.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: CRC Press.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147.

Schaie, K. W. (1959). Cross-sectional mnethods in the study of psychological aspects of aging. *J. Gerontol, 14*, 208-215.

Schaie, K. W. (2005). *Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study.* New York, NY: Oxford University Press.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*(1), 1.

Simpson, R. L., Lacava, P. G., & Graner, P. S. (2004). The No Child Lreft Behind Act: Challenges and Implications for Educators. *Intervention for Schools and Clinic, 40*(2), 67-75.

Skinner, B. F. (1938). The behavior of organisms: An experimental analysis. Oxford, England: Appleton-Century.

Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks. NCES 2009-004. *National Center for Education Statistics*.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: how to deal with missing data. *Journal of clinical epidemiology, 55*(4), 329-337.

Vygotsky, L. S. (1978). Mind and society: The development of higher mental processes: Cambridge, MA: Harvard University Press.

Warren, C. J. E. (1954). *Brown v. Board of Education*. United States Reports, 347, 483.

Watson, J. B. (1925). *Behaviorism*: Piscataway, NJ: Transaction Publishers.

West, J., Denton, K., & Germino-Hausken, E. (2000). *America's Kindergartners: Findings from the Early Childhood Longitufdinal Study, Kindergaten Class of 1998-99, Fall 1998.* Retrieved from http://eric.ed.gov/?id=ed438089

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance Within Longitidinal Structural Equation Models: Measuring the Same Contruct Across Time. *Child Development Perspectives, 4*(1), 10-18.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. . *Educational and Psychological Measurement, 49*, 587-602.

Yeung, W. J., & Conley, D. (2008). Black-White Achievement Gap and Family Wealth *Child Development, 79*(2), 303-324.

Appendices

Appendix 1 – Mplus syntax for data imputation

```
TITLE: Multiple data imputation with 10 imputed datasets

DATA:
   FILE IS ECLS-K.dat;

VARIABLE:
   NAMES ARE
   RACE ReadT1 ReadT2 ReadT4 ReadT5 ReadT6 ReadT7
   MathT1 MathT2 MathT4 MathT5 MathT6 MathT7
   NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
   SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
   CLSIZE COMPUT TREXP TRED SCHENRLS FLNCH RLNCH CLENV;

 USEVARIABLES ARE
   ALL;

   MISSING IS ALL (999);

DATA IMPUTATION:
   IMPUTE = race(c) readT1-mathT7 NUMSIB TYPFAMIL
   MOMED DADED PARED MOMSCR DADSCR SES POVRTY MUSEUM
   ZOO CHLBOO LIBRAR HOMECM CLSIZE COMPUT TREXP TRED
   SCHENRLS FLNCH RLNCH CLENV;    ! Race imputed as a categorical variable
   NDATASETS = 10;
   SAVE = ECLS-Kimp*.dat;

OUTPUT:
  STANDARDIZED;
```

Appendix 2 – Mplus syntax to test factorial invariance of indicator intercepts

```
TITLE: MIMIC invariance testing of indicator intercepts

 DATA:
   FILE IS ECLS-Kimplist.dat;
   TYPE=IMPUTATION;

 VARIABLE:
   NAMES ARE
     RACE ReadT1 ReadT2 ReadT4 ReadT5 ReadT6 ReadT7
     MathT1 MathT2 MathT4 MathT5 MathT6 MathT7
     NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
     SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
     CLSIZE COMPUT TREXP TRED SCHENRLS FLNCH RLNCH CLENV;

     USEVARIABLES ARE
      Race ReadT4 - ReadT6 MathT4-MathT6;
      MISSING = all (999);

 MODEL:
     Reading BY ReadT4* ReadT5 ReadT6;
     Math BY MathT4* MathT5 MathT6;
     Reading Math ON Race@0;  ! measurement invariance
     ReadT4 - ReadT6 MathT4-MathT6 ON Race; ! indicator difference
     reading@1;    ! fix reading variance to 1 to set the metric
     math@1;   ! fix math variance to 1 to set the metric

 OUTPUT:
  STANDARDIZED;
```

Appendix 3 – M*plus* syntax to test factorial invariance of factor means

```
TITLE: MIMIC invariance testing of factor means

DATA:
  FILE IS ECLS-Kimplist.dat;
  TYPE=IMPUTATION;

VARIABLE:
  NAMES ARE
    RACE ReadT1 ReadT2 ReadT4 ReadT5 ReadT6 ReadT7
    MathT1 MathT2 MathT4 MathT5 MathT6 MathT7
    NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
    SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
    CLSIZE COMPUT TREXP TRED SCHENRLS FLNCH RLNCH CLENV;

    USEVARIABLES ARE
     Race READT4-ReadT6 MathT4-MathT6;
     MISSING = all (999);

MODEL:
    Reading BY ReadT4-ReadT6;
    Math BY MathT4-MathT6;
    Reading Math ON Race; ! factor mean difference
    ReadT4-ReadT6 MathT4-MathT6 ON Race@0; ! measurement invariance
    reading@1;          ! fix reading variance to 1 to set the metric
    math@1;   ! fix math variance to 1 to set the metric

OUTPUT:
 STANDARDIZED;
```

Appendix 4 – M*plus* syntax for CFA without covariates

```
TITLE: CFA without covariates

 DATA:
  FILE IS ECLS-Kimplist.dat;
     TYPE = IMPUTATION;        ! 10 sets of imputed data

 VARIABLE:
  NAMES ARE
   RACE ReadT1 ReadT2 ReadT4 ReadT5 ReadT6 ReadT7
   MathT1 MathT2 MathT4 MathT5 MathT6 MathT7
   NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
   SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
   CLSIZE COMPUT TREXP TRED SCHENRLS FLNCH RLNCH CLENV;

   USEVARIABLES ARE
    READT4 READT5 READT6 MATHT4 MATHT5 MATHT6;
    MISSING = all (999);

 MODEL:
   reading BY readT4* readT5 readT6;     ! reading parameters are free to vary
   math BY mathT4* mathT5 mathT6;        ! math parameters are free to vary
   reading@1;          ! reading variance set at 1.0
   math@1;             ! math variance set at 1.0

 OUTPUT:
   STANDARDIZED;
```

Appendix 5 – Mplus syntax for CFA with race covariate

```
TITLE: CFA with race as a covariate

 DATA:
   FILE IS ECLS-Kimplist.dat; ! 10 sets of imputed data
        TYPE = IMPUTATION;

 VARIABLE:
   NAMES ARE
     RACE ReadT1 ReadT2 ReadT4 ReadT5 ReadT6 ReadT7
     MathT1 MathT2 MathT4 MathT5 MathT6 MathT7
     NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
     SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
     CLSIZE COMPUT TREXP TRED SCHENRLS FLNCH RLNCH CLENV;

     USEVARIABLES ARE
      Race READT4 READT5 READT6 MATHT4 MATHT5 MATHT6;
      MISSING = all (999);

 MODEL:
     reading BY readT4* readT5 readT6;
     math BY mathT4* mathT5 mathT6;
     reading math ON race;     !race added as a covariate
     reading@1;          ! reading variance set to 1.0
     math@1;             ! math variance set to 1.0

 OUTPUT:
     STANDARDIZED;
```

Appendix 6 – Mplus syntax for CFA, race, family, classroom and school covariates

```
TITLE: CFA with family, classroom and school covariates

 DATA:
  FILE IS ECLS-Kimplist.dat;        ! 10 sets of imputed data each with N = 15012
      TYPE = IMPUTATION;

 VARIABLE:
  NAMES ARE
    RACE ReadT1 ReadT2 ReadT4 ReadT5 ReadT6 ReadT7
    MathT1 MathT2 MathT4 MathT5 MathT6 MathT7
    NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
    SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
    CLSIZE COMPUT TREXP TRED SCHENRLS FLNCH RLNCH CLENV;

    USEVARIABLES ARE
     RACE READT4-READT6 MATHT4-MATHT6
     NUMSIB TYPFAMIL MOMED DADED PARED MOMSCR DADSCR
     SES POVRTY MUSEUM ZOO CHLBOO LIBRAR HOMECM
     CLSIZE COMPUT TREXP TRED
     SCHENRLS FLNCH RLNCH CLENV;
     MISSING = all (999);

 MODEL:
    Reading BY ReadT4-ReadT6;
    Math BY MathT4-MathT6;
    Reading math ON race;
    Reading math ON NUMSIB TYPFAMIL MOMED DADED
    PARED MOMSCR DADSCR SES POVRTY MUSEUM ZOO
    CHLBOO LIBRAR HOMECM; ! family characteristics
    Reading math ON CLSIZE COMPUT TREXP TRED;  ! classroom characteristics
    Reading math ON SCHENRLS FLNCH RLNCH CLENV;    ! school characteristics

 OUTPUT:
    STANDARDIZED;
```

Appendix 7 - Results of MIMIC model of reading, mathematics and covariates

Model parameter	Reading		Math	
	Unstandardized estimate/z-score	Standardized estimate/z-score	Unstandardized estimate/z-score	Standardized estimate/z-score
Race	-.402(-13.327)***	-.137(-13.481)***	-.580(-16.740)***	-.198(-16.870)***
Family covariates				
Number of siblings	-.059(-4.216)***	-.053(-4.209)***	-.015(-.969)ns	-.014(-.971)ns
Family type	-.000(-.012)ns	-.000(-.013)ns	-.005(.-329)ns	-.005(-.327)ns
Mother education	.071(4.507)***	.102(4.543)***	.077(3.577)***	.111(3.586)***
Father education	.093(7.384)***	.154(7.400)***	.084(5.195)***	.140(5.258)***
Parent education	-.054(-2.011)*	-.085(-2.001)*	-.047(-1.800)ns	-.073(-1.799)ns
Mother's occupational prestige	-.002(1.775)ns	-.025(-1.774)ns	-.001(-1.240)ns	-.015(-1.240)ns
Mother's occupational prestige	-.000(-.333)ns	-.006(-.330)ns	.000(-.167)ns	-.003(-.163)ns
Socioeconomic status measure	.320(4.688)***	.208(4.647)***	.289(5.491)***	.189(5.463)***
Level of poverty	.176(2.612)*	.048(2.639)**	.195(4.306)***	.053(4.328)***
Child visited museum	-.044(-1.510)ns	-.017(-1.505)ns	-.016(-.583)ns	-.007(-.582)ns
Child visited zoo	.089(3.239)**	.034(3.234)**	.080(3.234)**	.031(3.234)**
Number of books child has	.000(4.693)***	.044(4.689)***	.000(3.771)***	.036(3.781)***
Child visited library	-.109(-4.827)***	-.045(4.810)***	-.056(-2.258)*	-.023(-2.254)*
Child has home computer	-.284(-6.615)***	-.074(-6.641)***	-.250(-5.290)***	-.066(-5.323)***
Classroom covariates				
Class size	.001(.264)ns	.003(.264)ns	.003(.702)ns	.009(.704)ns
Computers in classroom	.011((1.616)ns	.017(.523)ns	.024(4.029)***	.038(4.085)***
Teacher's experience	.004(3.484)***	.037(3.519)***	.002(2.237)*	.021(2.253)*
Teacher's highest education	-.049(4.364)***	-.038(-4.368)***	-.038(-2.751)**	-.030(-2.765)**
Classroom climate (parceled)	-.184(-1.838)ns	-.018(-1.836)ns	-.436(-4.834)***	-.042(-4.826)***
School covariates				
School enrollment	.023(1.730)ns	.022(1.726)ns	.049(3.458)**	.048(3.719)**
Percent of free lunch eligible	-.005(6.825)***	-.102(-6.777)***	-.003(-3.729)***	-.059(-3.719)***
Percent of reduced lunch eligible	.003(.229)ns	.002(.228)ns	.019(1.463)ns	.017(1.417)ns

Note: Table shows unstandardized and standardized estimates and z-scores (in parentheses) of IVs. ns = not statistically significant at $p < .05$. * = statistically significant at $p < .05$. ** = statistically significant at $p < .01$. *** = statistically significant at $p < .001$. Classroom climate was parceled consisting of classroom's size and space, lighting, temperature, condition, ventilation, and noise level.

Appendix 8 - Univariate statistics and summary of estimated means from missing value analysis procedure

	Full sample			White			Black		
Variable	N	Mean/SD	% Missing	N	Mean/SD	% Missing	N	Mean/SD	%Missing
Race	15012	.21(.41)	0	11788	.00(.00)	0	3224	1.00(.00)	0
ReadT1	13262	35.60(9.96)	11.7	10418	36.46(10.31)	11.6	2844	32.43(7.79)	11.8
ReadT2	14016	46.81(13.89)	6.6	11059	48.00(14.27)	6.2	2957	42.36(11.35)	8.3
ReadT4	11777	78.68(23.84)	21.5	9420	81.22(23.90)	20.1	2357	68.53(20.71)	26.9
ReadT5	9938	130.12(27.47)	33.8	8094	134.25(26.24)	31.3	1844	111.96(25.30)	42.8
ReadT6	7739	153.43(25.53)	48.4	6465	157.07(23.78)	45.2	1274	134.92(26.04)	60.5
ReadT7	6605	173.89(26.32)	56.0	5663	177.20(23.71)	52.0	942	151.00(29.45)	70.8
MathT1	13262	26.87(9.12)	11.7	10418	28.08(9.32)	11.6	2844	22.27(6.75)	11.8
MathT2	14008	37.40(12.04)	6.7	11057	30.08(12.07)	6.2	2951	31.11(9.59)	8.5
MathT4	11776	62.97(18.32)	21.6	9420	65.65(18.20)	20.1	2356	52.26(14.48)	26.9
MathT5	9997	100.92(24.45)	33.4	8123	104.97(23.18)	31.1	1874	83.38(21.94)	41.9
MathT6	7745	125.57(24.23)	48.4	6470	129.39(22.28)	54.1	1275	106.16(24.45)	60.5
MathT7	6661	143.76(21.21)	55.6	5707	146.81(19.27)	51.6	954	125.52(23.03)	70.4
MOMED	6199	5.02(1.72)	58.7	5382	5.14(1.70)	54.3	817	4.21(1.57)	74.7
DADED	5154	5.12(1.95)	65.7	4742	5.19(1.95)	59.8	412	4.25(1.71)	87.2
PARED	6350	5.48(1.85)	57.7	5509	5.65(1.82)	53.3	841	4.39(1.62)	73.9
MOMSCR	6199	38.40(20.11)	58.7	5382	38.93(20.31)	54.3	817	34.94(18.38)	74.7
DADSCR	5154	43.34(14.72)	65.7	4742	43.96(14.47)	59.8	412	36.17(15.68)	87.2
SESL	6336	.10(.77)	57.8	5502	.19(.74)	53.3	834	-.47(.73)	74.1
POVRTY	6350	1.89(.32)	57.7	5509	1.93(.26)	53.3	841	1.62(.49)	73.9
MUSEUM	9548	1.66(.48)	36.4	7990	1.64(.48)	32.2	1558	1.72(.45)	51.7
ZOO	9550	1.70(.48)	36.4	7992	1.71(.46)	32.2	1558	1.68(.47)	51.7
CHLBOO	7563	120.19(182.72)	49.6	6409	130.76(189.82)	45.6	1154	61.49(121.56)	64.2
LIBRAR	7581	1.52(.50)	49.5	6421	1.52(.50)	45.5	1160	1.51(.50)	64.0
HOMECM	7586	1.11(.31)	49.5	6424	1.07(.26)	45.5	1162	1.32(.47)	64.0
CLSZ	8528	21.00(4.18)	43.2	7128	21.10(4.18)	39.5	1400	20.50(4.12)	56.6
COMPUT	8532	2.83(1.89)	43.2	7136	2.83(1.90)	39.5	1396	2.82(1.86)	56.7
TREXP	8501	15.35(10.18)	43.4	7101	15.6310.10)	39.8	1400	13.92(10.46)	56.6
TRHED	8484	2.20(.94)	43.5	7091	2.23(.92)	39.8	1393	2.07(1.01)	56.8
ENRLS	6398	3.71(1.17)	57.4	5585	3.69(1.18)	52.6	813	3.86(1.11)	74.8
FLNCH	8037	26.55(25.71)	46.5	6658	20.44(19.90)	43.5	1379	56.06(29.84)	57.2
RLNCH	8098	2.76(1.06)	46.1	6712	2.71(1.06)	43.1	1386	3.00(1.02)	57.0
CLEVN	13257	1.03(.12)	11.7	10793	1.03(.11)	8.4	2859	1.05(.14)	11.3

Note: 1). Continuous variables - ReadT4 – ReadT6: IRT reading scale score at time point 4, 5, and 6, MathematicsT4 – MathematicsT6: IRT mathematics scale score at time points 4, 5, and 6, MOMSCR: Mother's occupational prestige, DADSCR: Father's occupational prestige, SESL: Household income, CHLBOO: Number of books child has at home, ENRLS: Total school enrollment, FLNCH: Percent of free lunch eligible students, RLNCH: Percent of reduced lunch eligible students, CLSZ: Total number of students in class, TREXP: Number of years a school teacher; 2). Categorical variables - RACE: Race of child, MOMED: Mother's level of education, DADED: Father's level of education, PARED: Parent's level of education, POVRTY: Poverty level, LIBRAR: Visited the library, MUSEUM: Visited museum, ZOO: Visited zoo or aquarium, HOMECM: Has a computer at home for use, CLEVN: Classroom climate and environment, TRHED: Teacher's highest level of education achieved.

Appendix 9 – Description of variables

Variable	Description
ReadT1	Fall kindergarten IRT scale scores for reading at time point 1
ReadT2	Spring kindergarten IRT scale scores for reading at time point 2
ReadT4	Spring first grade IRT scale scores for reading at time point 4
ReadT5	Spring third grade IRT scale scores for reading at time point 5
ReadT6	Spring fifth grade IRT scale scores for reading at time point 6
ReadT7	Spring eighth grade IRT scale scores for reading at time point 7
MathT1	Fall kindergarten IRT scale scores for mathematics at time point 1
MathT2	Spring kindergarten IRT scale scores for mathematic at time point 2
MathT4	Spring first grade IRT scale scores for mathematics at time point 4
MathT5	Spring third grade IRT scale scores for mathematics at time point 5
MathT6	Spring fifth grade IRT scale scores for mathematic at time point 6
MathT7	Spring eighth grade IRT scale scores for mathematics at time point 7
NUMSIB	Number of siblings living at home
TYPFAMIL	Type of family, single- or two-parent family
MOMED	Mother's highest level of education
DADED	Father's highest level of education
PARED	Parent highest level of education
MOMSCR	Mother's occupation prestige score
DADSCR	Father's occupation prestige score
SES	Socioeconomic status of family
POVRTY	Level of poverty
MUSEUM	Child's visit to the museum
ZOO	Child's visit to the zoo
CHLBOO	Number of books child has at home
LIBRAR	Child's visit to the library
HOMECM	Home computer child uses
CLSIZE	Class size
COMPUT	Number of computers in teacher's classroom
TREXP	Number of years teacher has been teaching
TRED	Highest level of education achieved by teacher
CLENV	Classroom climate and environment (parceled)
SCHENRLS	Total school enrolment
FLNCH	Percent of students eligible for free lunch
RLNCH	Percent of students eligible for reduced lunch